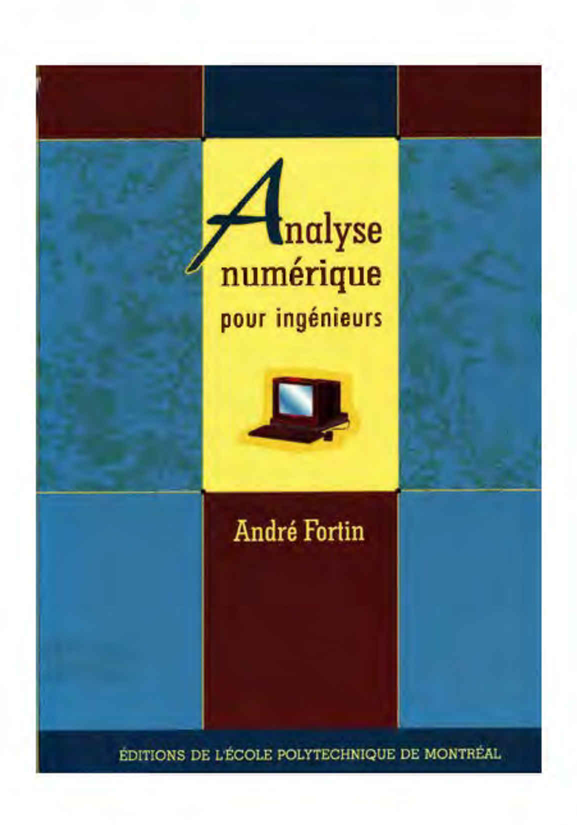
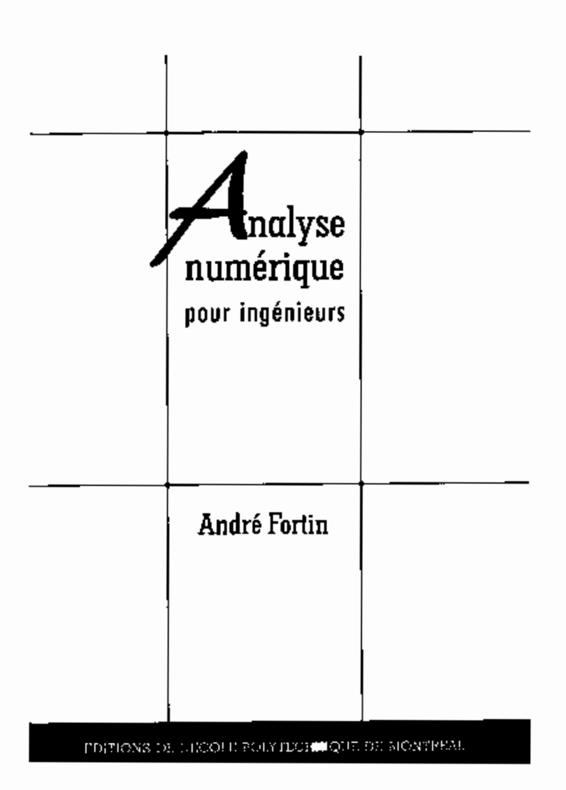

Analyse
numérique
pour ingénieurs
André Fortin
ÉDITIONS DE L'ÉCOLE POLYTECHNIQUE DE MONTRÉAL


Analyse
numérique
pour ingénieurs
André Fortin
EDITIONS DE L'ECOLE POLYTECHNIQUE DE MONTREAL

Avant-propos

Ce manuel reflète mon expérience comme professeur et comme coordonnateur du cours d'analyse numérique 215 offert aux étudiants de l'École Polytechnique de Montréal. Chaque année, environ 500 étudiants suivent ce cours qui propose un survol des principales méthodes numériques élémentaires et couvre plus particulièrement les sujets suivants:

- analyse d'erreurs;
- racines d'une équation algébrique;
- systèmes d'équations linéaires et non linéaires;
- interpolation, différentiation et intégration numériques;
- équations différentielles ordinaires.

La première tâche du coordonnateur consiste à trouver un livre qui convienne parfaitement à la matière fixée par les exigences du programme d'enseignement. On trouve sur le marché de nombreux livres qui traitent de l'analyse numérique. Cette branche des mathématiques appliquées connaît en effet un essor considérable depuis plusieurs années et à peu près toutes les facultés de génie offrent au moins un cours d'introduction à l'analyse numérique, suivi très souvent d'un second cours plus avancé. On peut mettre la main sur nombre de bons manuels conçus aux États-Unis et donc publiés en anglais. Du côté québécois, comme on peut s'y attendre, la production est mince en raison de l'étroitesse du marché. Enfin, les ouvrages en provenance d'Europe ne conviennent pas à nos besoins en raison de leur contenu mathématique trop avancé.

Cet état de fait m'a conduit à tenter l'expérience de rassembler mes notes de cours et de publier ce manuel. Le lecteur y trouvera une introduction à l'analyse numérique élémentaire motivée par les applications en ingénierie.

Il m'a paru intéressant d'aborder du même coup un domaine en pleine expansion que j'ai eu l'occasion d'explorer pendant mes activités de recherche, à savoir les systèmes dynamiques. Il ne s'agit ici que d'illustrer la contribution des méthodes numériques au développement de ce champ d'activités. J'attire tout au plus l'attention sur quelques comportements des systèmes dynamiques qui peuvent être observés à l'aide de méthodes numériques élémentaires.

L'approche pédagogique de ce manuel repose sur une compréhension profonde des méthodes plutôt que sur l'aspect calculatoire. Cela signifie que les exemples choisis cherchent avant tout à illustrer différents aspects des méthodes et à souligner leurs avantages et leurs inconvénients. Cette approche est justifiée en partie par le fait que de plus en plus d'ingénieurs utilisent des outils logiciels commerciaux. L'objectif de ce manuel est donc de faire des étudiants des utilisateurs intelligents, en ce sens qu'ils sauront exactement à quoi s'attendre de chaque méthode et qu'ils seront en mesure de valider leurs résultats.

En terminant, j'aimerais remercier toutes les personnes qui ont contribué à la réalisation de ce manuel. En particulier, Mme Carole Burney-Vincent et M. Gilles Savard du département de mathématiques et de génie industriel ont patiemment lu et commenté plusieurs chapitres. M. Robert Roy a accepté de se servir d'une version préliminaire pour le cours d'analyse numérique donné pendant l'été de 1994; ses remarques et ses suggestions ont été très utiles. J'ai également pu compter sur la collaboration de toute l'équipe professorale du cours d'analyse numérique. Il me faut également souligner la participation du Service Pédagogique de l'École Polytechnique de Montréal.

Enfin, je ne peux passer sous silence l'appui inconditionnel de mon épouse Marie et de mes fils Michel et Jean-Philippe qui ont dû, entre autres choses, subir mes absences fréquentes lors de la rédaction et de la mise en pages finale de cet ouvrage. Je tiens de plus à souligner que la page couverture de ce manuel est inspirée des dessins de Michel et Jean-Philippe.

Veuillez tous trouver ici l'expression de ma plus profonde reconnaissance.

Table des matières

Chapitre 1

Analyse d'erreurs

1.1 Introduction

Les cours traditionnels de mathématiques nous familiarisent avec des théories et des méthodes qui permettent de résoudre de façon *analytique* un certain nombre de problèmes. C'est le cas notamment des techniques d'intégration et de résolution d'équations algébriques ou différentielles. Bien qu'on puisse proposer plusieurs méthodes pour résoudre un problème donné, celles-ci conduisent à un même résultat, précis et unique.

C'est ici que l'analyse numérique se distingue des autres champs plus classiques des mathématiques. En effet, pour un problème donné, il est possible d'utiliser plusieurs techniques de résolution qui résultent en différents algorithmes. Ces algorithmes dépendent de certains paramètres qui influent sur la précision du résultat. De plus, on utilise en cours de calcul des approximations plus ou moins précises. Par exemple, on peut remplacer une dérivée par une différence finie de façon à transformer une équation différentielle en une équation algébrique. Le résultat final et son degré de précision dépendent des choix que l'on fait.

Une partie importante de l'analyse numérique consiste donc à contenir les effets des erreurs ainsi introduites, qui proviennent de trois sources principales:

- les erreurs de modélisation;
- les erreurs de représentation sur ordinateur;
- les erreurs de troncature.

Les erreurs de modélisation, comme leur nom l'indique, proviennent de l'étape de mathématisation du phénomène physique auquel on s'intéresse. Cette étape consiste à faire ressortir les causes les plus déterminantes du phénomène observé et à les mettre sous forme d'équations (différentielles le plus souvent). Si le phénomène observé est très complexe, il faut simplifier et négliger ses composantes qui paraissent moins importantes ou qui rendent la résolution numérique trop difficile. C'est ce que l'on appelle les *erreurs de modélisation*.

La seconde catégorie d'erreurs est liée à l'utilisation de l'ordinateur. En effet, la représentation sur ordinateur (généralement binaire) des nombres introduit souvent des erreurs. Même infimes au départ, ces erreurs peuvent s'accumuler lorsqu'on effectue un très grand nombre d'opérations. Par exemple, la fraction $1/3$ n'a pas de représentation binaire exacte, car elle ne possède pas de représentation décimale finie. Ces erreurs se propagent au fil des calculs et peuvent compromettre la précision des résultats.

Enfin, les *erreurs de troncature* proviennent principalement de l'utilisation du développement de Taylor, qui permet par exemple de remplacer une équation différentielle par une équation algébrique. Le développement de Taylor est le principal outil mathématique du numéricien. Il est donc primordial d'en maîtriser l'énoncé et ses conséquences.

Ce chapitre traite donc principalement d'erreurs numériques, et non des inévitables erreurs de programmation qui font, hélas, partie du quotidien du numéricien. Il devrait permettre au lecteur de mieux gérer les erreurs au sein des processus numériques afin d'être en mesure de mieux interpréter les résultats.

Tout d'abord, un peu de terminologie est nécessaire pour éventuellement quantifier les erreurs.

Définition 1.1
Soit x, un nombre, et x^*, une approximation de ce nombre. L'*erreur absolue* est définie par:

$$\Delta x = |x - x^*| \tag{1.1}$$

Définition 1.2
Soit x, un nombre, et x^*, une approximation de ce nombre. L'*erreur relative* est définie par:

$$E_r(x) = \frac{|x - x^*|}{|x|} = \frac{|\Delta x|}{|x|} \tag{1.2}$$

De plus, en multipliant par 100 %, on obtient l'*erreur relative en pourcentage.*

En pratique, il est difficile d'évaluer les erreurs absolue et relative, car on ne connaît généralement pas la valeur exacte de x et on n'a que x^*. Dans le cas de quantités mesurées dont on ne connaît que la valeur approximative, il est impossible de calculer l'erreur absolue; on dispose en revanche d'une borne supérieure pour cette erreur qui dépend de la précision des instruments de mesure utilisés. Cette borne est quand même appelée erreur absolue, alors qu'en fait on a:

$$|x - x^*| \leq \Delta x$$

ce qui peut également s'écrire:

$$x^* - \Delta x \leq x \leq x^* + \Delta x \tag{1.3}$$

et que l'on note parfois $x = x^* \pm \Delta x$. On peut interpréter ce résultat en disant que l'on a estimé la valeur exacte x à partir de x^* avec une incertitude de Δx de part et d'autre.

L'erreur absolue donne une mesure quantitative de l'erreur commise et l'erreur relative en mesure l'importance. Par exemple, si on fait usage d'un chronomètre dont la précision est de l'ordre du dixième de seconde, l'erreur absolue est bornée par 0,1 s. Mais est-ce une erreur importante? Dans le contexte d'un marathon d'une durée de 2 h 20 min, l'erreur relative liée au chronométrage est très faible:

$$\frac{0{,}1}{2 \times 60 \times 60 + 20 \times 60} = 0{,}000\,0119$$

et ne devrait pas avoir de conséquence sur le classement des coureurs. Par contre, s'il s'agit d'une course de 100 m d'une durée d'environ 10 s, l'erreur relative est beaucoup plus importante:

$$\frac{0{,}1}{10{,}0} = 0{,}01$$

soit 1 % du temps de course. Avec une telle erreur, on ne pourra vraisemblablement pas faire la différence entre le premier et le dernier coureur.

Cela nous amène à parler de précision et de *chiffres significatifs* au sens de la définition suivante.

Définition 1.3
Si l'erreur absolue vérifie

$$\Delta x \leq 0{,}5 \times 10^m$$

alors le chiffre correspondant à la m^e puissance de 10 est dit *significatif* et tous ceux à sa gauche (correspondant aux puissances de 10 supérieures à m) le sont aussi.

Exemple 1.1
On obtient une approximation de π ($x = \pi$) au moyen de la quantité 22/7 ($x^* = 22/7 = 3{,}142\,857 \cdots$). On en conclut que:

$$\Delta x = |\pi - \frac{22}{7}| = 0{,}001\,26 \cdots$$

Puisque l'erreur absolue est plus petite que $0{,}5 \times 10^{-2}$, le chiffre des centièmes est significatif et on a en tout 3 chiffres significatifs (3,14).

• • • • •

Exemple 1.2
Si on retient 3,1416 comme approximation de π, on a:

$$\Delta x = |\pi - 3{,}1416| \simeq 0{,}73 \times 10^{-5}$$

et l'erreur absolue est inférieure à $0{,}5 \times 10^{-4}$. Le chiffre correspondant à cette puissance de 10 (6) est significatif au sens de la définition, ainsi que tous les chiffres situés à sa gauche. Il est à remarquer que le chiffre 6 dans 3,1416 est significatif même si la quatrième décimale de π est un 5 ($3{,}141\,59 \cdots$).

L'approximation 3,1416 possède donc 5 chiffres significatifs.

• • • • •

Remarque 1.1
Inversement, si un nombre est donné avec n chiffres significatifs, cela signifie que l'erreur absolue est inférieure à 0,5 fois la puissance de 10 correspondant au dernier chiffre significatif. □

Exemple 1.3
On a mesuré le poids d'une personne et trouvé 90,567 kg. On vous assure que l'appareil utilisé est suffisamment précis pour que tous les chiffres fournis soient significatifs. D'après la définition, puisque le dernier chiffre significatif correspond aux millièmes (milligrammes), cela signifie que:

$$\Delta x \leq 0{,}5 \times 10^{-3} \text{ kg}$$

En pratique, on conclut que:

$$\Delta x = 0{,}5 \times 10^{-3} \text{ kg}$$

• • • • •

1.2 Erreurs de modélisation

La première étape de la résolution d'un problème, et peut-être la plus délicate, consiste à modéliser le phénomène observé. Il s'agit en gros d'identifier tous les facteurs internes et externes qui influent (ou que l'on soupçonne d'influer) sur les résultats. Dans le cas d'un phénomène physique, on fait l'inventaire des forces en présence: gravitationnelle, de friction, électrique, etc. On a par la suite recours aux lois de conservation de la masse, de l'énergie, de la quantité de mouvement et à d'autres principes mathématiques pour traduire l'influence de ces différents facteurs sous forme d'équations. Le plus souvent, on obtient des équations différentielles ou aux dérivées partielles.

L'effort de modélisation produit en général des systèmes d'équations complexes qui comprennent un grand nombre de variables inconnues. Pour réussir à les résoudre, il faut simplifier certaines composantes et négliger les moins importantes. On fait alors une première erreur de modélisation.

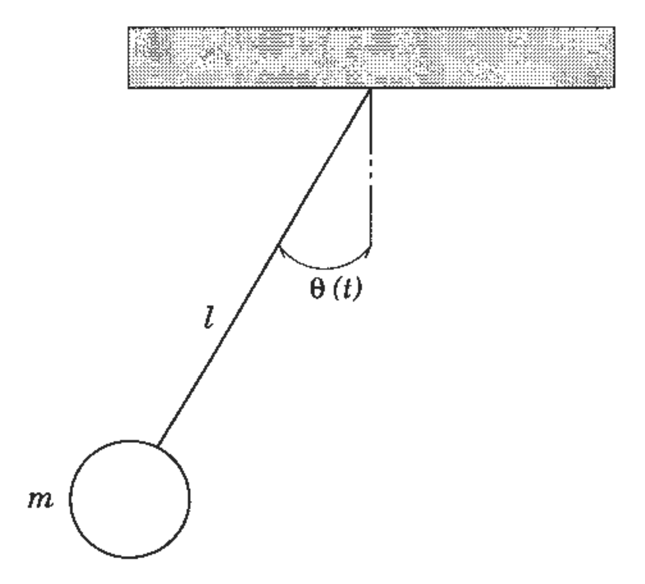


Figure 1.1: Problème du pendule

De plus, même bien décomposé, un phénomène physique peut être difficile à mettre sous forme d'équations. On introduit alors un modèle qui décrit au mieux son influence, mais qui demeure une approximation de la réalité. On commet alors une deuxième erreur de modélisation. Illustrons cette démarche à l'aide d'un exemple.

Exemple 1.4
Le problème du pendule est connu depuis très longtemps (voir par exemple Simmons, réf. [20]). Une masse m est suspendue à une corde de longueur l (voir la figure 1.1). Au temps $t = 0$, on suppose que l'angle $\theta(0)$ entre la corde et la verticale est θ_0 et que sa vitesse angulaire $\theta'(0)$ est θ'_0. Les forces en présence sont la gravité agissant sur la masse et la corde, d'une part, et la friction de l'air agissant sur tout le système, d'autre part.

Suivant la deuxième loi de Newton, la force due à l'accélération tangentielle $ml\theta''(t)$ est équilibrée par la composante tangentielle de l'accélération gravitationnelle $mg\sin(\theta(t))$ et par la force de friction F_f qui s'oppose au mouvement. On a alors:

$$ml\theta''(t) = -mg\sin(\theta(t)) - F_f$$

Pour connaître comment se comporte la force de friction F_f en fonction de $\theta(t)$, il faut recourir à des mesures expérimentales, qui démontrent que la friction est à peu près proportionnelle à la vitesse $l\theta'(t)$ avec un coefficient de proportionnalité c_f. Il est important de remarquer que cette loi est approximative et que le coefficient c_f est obtenu avec une précision limitée. Tout cela entraîne des erreurs de modélisation.

On obtient une équation différentielle du second ordre:

$$\theta''(t) = -\frac{c_f\theta'(t)}{m} - \frac{g\sin(\theta(t))}{l} \tag{1.4}$$
$$\theta(0) = \theta_0 \tag{1.5}$$
$$\theta'(0) = \theta'_0 \tag{1.6}$$

L'équation différentielle 1.4 est non linéaire et on démontre qu'elle ne possède pas de solution analytique simple.

Une brève analyse montre qu'il est raisonnable de négliger la friction de l'air, car les vitesses de mouvement sont faibles. Il s'agit encore d'une erreur de modélisation, qui paraît acceptable à cette étape de l'analyse. Si les résultats s'avèrent insatisfaisants, il faudra revenir en arrière et identifier, parmi les forces négligées, celle qui était la plus importante. Une analyse adimensionnelle est souvent nécessaire pour y arriver.

Même en négligeant la friction, ce qui revient à poser $c_f = 0$, l'équation résultante ne possède toujours pas de solution simple. En effet, sa résolution fait intervenir les intégrales elliptiques (Simmons, réf. [20]) qui ne peuvent s'exprimer en fonctions élémentaires. Puisque tel est le but, on doit encore simplifier le problème. On peut par exemple supposer que les angles sont petits et que:

$$\sin(\theta(t)) \approx \theta(t)$$

Il en résulte le problème simplifié suivant:

$$\theta''(t) = -\frac{g\,\theta(t)}{l} \tag{1.7}$$
$$\theta(0) = \theta_0 \tag{1.8}$$
$$\theta'(0) = \theta'_0 \tag{1.9}$$

L'équation 1.7 possède la solution périodique classique:

$$\theta(t) = A\cos(\omega t) + B\sin(\omega t) \tag{1.10}$$

où $\omega^2 = g/l$ et les constantes A et B sont déterminées par les conditions initiales ($A = \theta_0$, $B = \theta'_0/\omega$).

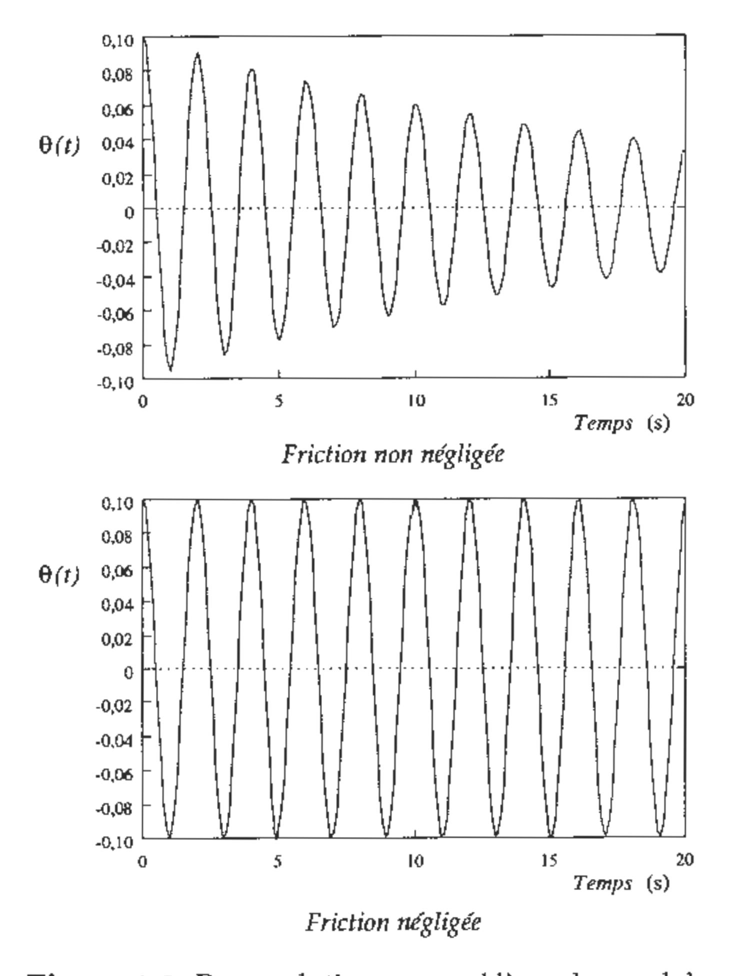


Figure 1.2: Deux solutions au problème du pendule

La figure 1.2 permet la comparaison entre les solutions des équations différentielles 1.4 et 1.7 dans le cas où $\theta_0 = 0{,}1$ et $\theta'_0 = 0$. L'équation 1.10 est la solution analytique de l'équation différentielle 1.7, alors que l'équation différentielle 1.4 ne possède qu'une solution numérique (nous la reverrons au chapitre 7). On remarque immédiatement que la solution numérique (où la friction n'est pas négligée) prévoit l'amortissement du mouvement avec le temps, tandis que la solution analytique reste parfaitement périodique. La solution numérique est de toute évidence plus près de ce que l'on observe pour un pendule.

• • • • •

Remarque 1.2
Les erreurs de modélisation, quoique très importantes, ne font pas partie de la matière de ce livre. Nous faisons l'hypothèse que les problèmes étudiés sont bien modélisés, bien que ce ne soit pas toujonrs le cas en pratique. □

1.3 Représentation des nombres sur ordinateur

Un ordinateur ne peut traiter les nombres de la même manière que l'être humain. Il doit d'abord les représenter dans un système qui permette l'exécution efficace des diverses opérations. Cela peut entraîner des erreurs de *représentation sur ordinateur*, qui sont inévitables et dont il est souhaitable de comprendre l'origine afin de mieux en maîtriser les effets. Cette section présente les principaux éléments d'un modèle de représentation des nombres sur ordinateur. Ce modèle s'avère utile pour comprendre le fonctionnement type des ordinateurs; il ne peut bien entendu tenir compte des caractéristiques sans cesse changeantes des ordinateurs modernes.

La structure interne de la plupart des ordinateurs s'appuie sur le système binaire. L'unité d'information ou *bit* prend la valeur 0 ou 1. Évidemment, très peu d'information peut être accumulée au moyen d'un seul bit. On regroupe alors les bits en *mots* de longueur variable (les longueurs de 8, de 16, de 32 ou de 64 bits sont les plus courantes). Les nombres, entiers et réels, sont représentés de cette manière, bien que leur mode précis de représentation dépende du fabricant.

1.3.1 Représentation des entiers signés

Nous traitons de trois formes parmi les plus courantes de représentation des entiers sur ordinateur. Plusieurs autres variantes existent, et il importe de connaître celle qui est utilisée par l'ordinateur afin de pouvoir s'y retrouver.

Pour transformer un entier positif N dans sa représentation binaire habituelle, il faut déterminer les a_i tels que:

$$(N)_{10} = (a_{n-1}a_{n-2}a_{n-3}\cdots a_2a_1a_0)_2$$

ou encore

$$N = a_{n-1}\times 2^{n-1} + a_{n-2}\times 2^{n-2} + a_{n-3}\times 2^{n-3} + \cdots + a_2\times 2^2 + a_1\times 2^1 + a_0\times 2^0$$

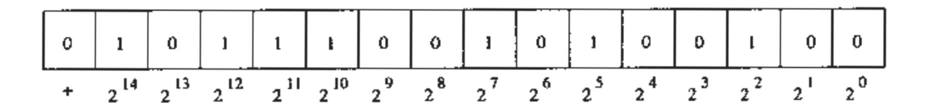


Figure 1.3: Représentation signe et grandeur sur 16 bits d'un entier

Dans ce qui précède, le sous-indice indique la base utilisée. On obtient la valeur des a_i en suivant la démarche suivante: en divisant N par 2, on obtient a_0 (le reste de la division) plus un entier; on refait le même raisonnement avec la partie entière de $N/2$ (en négligeant la partie fractionnaire ou reste) pour obtenir a_1; on continue ainsi jusqu'à ce que la partie entière soit nulle.

Exemple 1.5
Si $N = (1000)_{10}$, on a:

$$\begin{array}{rclll} 1000/2 & = & 500 & \text{reste } 0 & \text{c.-à-d.} \quad a_0 = 0 \\ 500/2 & = & 250 & \text{reste } 0 & \text{c.-à-d.} \quad a_1 = 0 \\ 250/2 & = & 125 & \text{reste } 0 & \text{c.-à-d.} \quad a_2 = 0 \\ 125/2 & = & 62 & \text{reste } 1 & \text{c.-à-d.} \quad a_3 = 1 \\ 62/2 & = & 31 & \text{reste } 0 & \text{c.-à-d.} \quad a_4 = 0 \\ 31/2 & = & 15 & \text{reste } 1 & \text{c.-à-d.} \quad a_5 = 1 \\ 15/2 & = & 7 & \text{reste } 1 & \text{c.-à-d.} \quad a_6 = 1 \\ 7/2 & = & 3 & \text{reste } 1 & \text{c.-à-d.} \quad a_7 = 1 \\ 3/2 & = & 1 & \text{reste } 1 & \text{c.-à-d.} \quad a_8 = 1 \\ 1/2 & = & 0 & \text{reste } 1 & \text{c.-à-d.} \quad a_9 = 1 \end{array}$$

Ainsi, l'entier décimal 1000 s'écrit 11 1110 1000 en base 2.

● ● ● ● ●

Voyons maintenant quelques variantes de représentation binaire des entiers signés: la représentation signe et grandeur, la représentation en complément à 2 et la représentation par excès. Ces variantes présentent toutes un certain nombre d'avantages et d'inconvénients au regard de l'exécution des opérations arithmétiques élémentaires.

Représentation signe et grandeur

Pour illustrer la représentation signe et grandeur, il suffit de prendre un exemple. Considérons un mot de 16 bits tel qu'illustré à la figure 1.3. Dans

cette représentation, un bit (le premier) est consacré au signe:

$$\begin{array}{ll} 0 & \text{pour un entier positif} \\ 1 & \text{pour un entier négatif} \end{array}$$

Les autres bits peuvent alors servir à la représentation de l'entier lui-même. Ainsi:

$$0101\,1100\,1010\,0100 =$$
$$+\ \left(1 \times 2^{14} + 0 \times 2^{13} + 1 \times 2^{12} + \cdots + 1 \times 2^2 + 0 \times 2^1 + 0 \times 2^0\right)$$

est la représentation binaire du nombre décimal 23 716. Le premier bit (0) indique un entier positif. Le plus grand entier représentable dans ce cas est bien sûr 0111 1111 1111 1111 qui représente 32 767, c'est-à-dire $2^{15} - 1$ en base 10. Il est à remarquer que le nombre 0 peut être représenté de deux façons, à savoir:

$$\begin{array}{rcl} +0 & = & 0000\,0000\,0000\,0000 \\ -0 & = & 1000\,0000\,0000\,0000 \end{array}$$

Un mot de 16 bits permet d'exprimer tous les entiers compris entre $-32\,767$ et $+32\,767$. Si un calcul sur des nombres entiers résulte en un entier supérieur à 32 767, le compilateur enverra un message d'erreur indiquant un *débordement* (*overflow*).

Remarque 1.3
Dans la représentation signe et grandeur et également dans les représentations qui suivent, nous utilisons la convention que le premier bit est celui situé le plus à gauche. En informatique, on utilise plus souvent une numérotation des bits allant de 0 jusqu'à $n - 1$ en commençant par le bit le plus à droite dit *le moins significatif.* □

Représentation en complément à 2

La représentation en complément à 2 est très fréquemment utilisée. Si on dispose de n bits pour exprimer l'entier N, on pose:

$$N = -a_{n-1} \times 2^{n-1} + a_{n-2} \times 2^{n-2} + a_{n-3} \times 2^{n-3} + \cdots + a_2 \times 2^2 + a_1 \times 2^1 + a_0 \times 2^0$$

Il faut remarquer le signe négatif devant le terme a_{n-1}. On constate facilement que tous les entiers positifs vérifient:

$$a_{n-1} = 0$$

Les entiers positifs sont donc représentés par 0 suivi de leur expression binaire habituelle en $(n-1)$ bits. Pour obtenir la représentation d'un nombre négatif, il suffit de lui ajouter 2^{n-1} et de transformer le résultat en forme binaire.

Exemple 1.6
La représentation sur 4 bits de 0101 vaut:

$$-0 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$$

soit 5 en forme décimale. Par contre, 1101 vaut:

$$-1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$$

c'est-à-dire $-8 + 5 = -3$. Inversement, la représentation binaire de -6 sera 1 suivi de la représentation sur 3 bits de:

$$-6 + 2^3 = 2$$

qui est 010. On aura donc $(-6)_{10} = (1010)_2$ dans la représentation en complément à 2.

●●●●●

Représentation par excès

Illustrons la représentation par excès en prenant pour exemple un mot de 4 bits ($n = 4$). On peut alors représenter au plus 2^4 (2^n dans le cas général) entiers différents, y compris les entiers négatifs. Si on veut exprimer un entier décimal N, il suffit de lui ajouter un excès d et de donner le résultat sous forme binaire. Inversement, si on a la représentation binaire par excès d'un entier, il suffit de calculer sa valeur en base 10 et de soustraire d pour obtenir l'entier recherché.

La représentation par excès a l'avantage d'ordonner la représentation binaire en assignant à 0000 le plus petit entier décimal représentable, à savoir $-d$. En général, la valeur de d est 2^{n-1} (2^3 sur 4 bits). Ainsi, avec 4 bits et $d = 2^3$, on obtient:

Forme binaire	Forme décimale
0000	-8
0001	-7
$\vdots$	$\vdots$
1110	$+6$
1111	$+7$

Pour obtenir ce tableau, il suffit de remarquer que, par exemple, 1111 vaut 15 en décimal, auquel on soustrait 2^3 pour obtenir 7.

Exemple 1.7
Soit un mot de 8 bits et un excès $d = 2^7 = 128$. Pour représenter $(-100)_{10}$, il suffit de lui ajouter 128, ce qui donne 28, et d'exprimer le résultat sur 8 bits, soit 0001 1100.

● ● ● ● ●

Remarque 1.4
Il existe d'autres représentations des entiers signés, comme la *représentation en complément à 1*. Notre but n'étant pas d'en donner une liste exhaustive, nous nous limitons à celles qui ont déjà été présentées. □

1.3.2 Représentation des nombres réels

La tâche de représentation est plus complexe dans le cas des nombres réels. En effet, dans le système décimal, nous avons l'habitude de représenter un nombre x sous la forme:

$$x = m \times 10^l$$

où m est la *mantisse*, l est l'*exposant* et 10 est la *base*. De façon générale, selon une base b quelconque, on peut écrire:

$$x = m \times b^l$$

La forme générale de la mantisse est la suivante:

$$m = 0{,}d_1 d_2 d_3 \cdots d_n$$

ce qui signifie que:

$$m = d_1 \times b^{-1} + d_2 \times b^{-2} + d_3 \times b^{-3} + \cdots + d_n \times b^{-n}$$

où n est le nombre de chiffres de la mantisse. Les d_i vérifient:

$$1 \leq d_1 \leq (b-1) \quad (1.11)$$
$$0 \leq d_i \leq (b-1) \text{ pour } i = 2, 3, \cdots, n \quad (1.12)$$

La première inégalité signifie que la mantisse est *normalisée*, c'est-à-dire que son premier chiffre est toujours différent de 0. La normalisation permet d'assurer l'unicité de la représentation et d'éviter les ambiguïtés entre:

$$0{,}1020 \times 10^2 \text{ et } 0{,}0102 \times 10^3$$

pour représenter 10,2. La dernière expression n'est jamais retenue. Dans cet exemple, on a utilisé la base $b = 10$ et $n = 4$ chiffres dans la mantisse. Ainsi, la mantisse satisfait toujours:

$$\frac{1}{b} \leq m < 1$$

car la plus petite mantisse possible est 0,1000, qui vaut $1/b$.

Ces considérations servent de lignes directrices pour la représentation d'un nombre réel sur ordinateur. Dans le cas qui nous intéresse, la base sera généralement 2 ($b = 2$). Il faut donc trouver un moyen de représenter la mantisse (une fraction), l'exposant (un entier signé) et le signe de ce nombre.

Remarque 1.5
Les calculatrices de poche se distinguent des ordinateurs principalement par le fait qu'elles utilisent la base 10 ($b = 10$) et une mantisse d'une longueur d'environ 10 ($n = 10$). L'exposant l varie généralement entre -100 et 100.□

Considérons à titre d'exemple un mot de 8 bits tel que l'illustre la figure 1.4. Il est entendu qu'en pratique les mots sont beaucoup plus grands. Le premier bit donne le signe du nombre. Il reste donc 7 bits pour représenter la mantisse et l'exposant. On retient 3 bits pour l'exposant et les 4 bits suivants pour la mantisse. Il est à noter que cette dernière est normalisée et que son premier bit est toujours 1. Il faut ensuite déterminer quelle est la représentation utilisée pour l'exposant. Par exemple, si l'exposant est exprimé suivant la *représentation signe et grandeur*, alors 0101 1011 se décompose de la façon suivante:

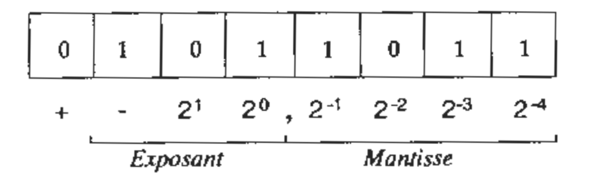


Figure 1.4: Représentation signe et grandeur sur 8 bits d'un réel

Signe	Exposant	Mantisse
0	101	1011

Ce nombre est donc positif. L'exposant est négatif et vaut:

$$-(0 \times 2^1 + 1 \times 2^0) = -1$$

et la mantisse est:

$$1 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4} = 0{,}6875$$

Ainsi, 0101 1011 représente le nombre décimal:

$$0{,}6875 \times 2^{-1} = 0{,}343\,75$$

Par contre, si l'exposant est exprimé suivant la représentation en *complément à 2*, la valeur binaire 101 qui représente l'exposant signifie alors -3 et $(0101\ 1011)_2$ vaut 0,085 9375 en base 10.

Conversion d'une fraction décimale en valeur binaire

La méthode de conversion d'une fraction décimale en valeur binaire est similaire à celle que l'on utilise dans le cas des entiers. Soit f, une fraction décimale comprise entre 0 et 1. Il faut donc trouver les d_i tels que:

$$(f)_{10} = (0{,}d_1 d_2 d_3 \cdots)_2$$

ou encore

$$f = d_1 \times 2^{-1} + d_2 \times 2^{-2} + d_3 \times 2^{-3} + \cdots$$

Si on multiplie f par 2, on obtient d_1 plus une fraction. En appliquant le même raisonnement à $(2f - d_1)$, on obtient d_2. On poursuit ainsi jusqu'à ce que la partie fractionnaire soit nulle ou que l'on ait atteint le nombre maximal de chiffres de la mantisse.

Exemple 1.8
Si $f = 0{,}0625$, on a:

$$\begin{array}{rcll} 0{,}0625 \times 2 & = & 0{,}1250 & \text{c.-à-d.} \quad d_1 = 0 \\ 0{,}1250 \times 2 & = & 0{,}2500 & \text{c.-à-d.} \quad d_2 = 0 \\ 0{,}2500 \times 2 & = & 0{,}5000 & \text{c.-à-d.} \quad d_3 = 0 \\ 0{,}5000 \times 2 & = & 1{,}0000 & \text{c.-à-d.} \quad d_4 = 1 \end{array}$$

ce qui signifie que $(0{,}0625)_{10} = (0{,}0001)_2$.

● ● ● ● ●

Exemple 1.9
Si $f = 1/3$, on a:

$$\begin{array}{rcll} 1/3 \times 2 & = & 0 + 2/3 & \text{c.-à-d.} \quad d_1 = 0 \\ 2/3 \times 2 & = & 1 + 1/3 & \text{c.-à-d.} \quad d_2 = 1 \\ 1/3 \times 2 & = & 0 + 2/3 & \text{c.-à-d.} \quad d_3 = 0 \\ 2/3 \times 2 & = & 1 + 1/3 & \text{c.-à-d.} \quad d_4 = 1 \\ & & & \\ \vdots & & \vdots & \qquad\qquad \vdots \end{array}$$

On peut poursuivre la conversion à l'infini et démontrer que:

$$\frac{1}{3} = (0{,}010\,101\cdots)_2$$

En pratique, puisqu'on n'utilise qu'un nombre fini de chiffres dans la mantisse, il faudra s'arrêter après n bits.

● ● ● ● ●

Norme IEEE

L'Institute for Electrical and Electronic Engineers (IEEE) s'efforce de rendre aussi uniformes que possible les représentations sur ordinateur. Il propose nne représentation des nombres réels en *simple précision* sur 32 bits et en *double précision* sur 64 bits (convention IEEE-754). Les représentations

en simple et double précision sont construites comme suit. Le premier bit témoigne du signe du nombre, les 8 bits suivants (11 en double précision) déterminent l'exposant avec un excès de 127 ou $2^{8-1}-1$ (1023 ou $2^{11-1}-1$ en double précision) et les 23 derniers bits (52 en double précision) sont pour la mantisse normalisée. Puisque l'on normalise la mantisse, le premier bit est toujours 1 et il n'est pas nécessaire de le garder en mémoire. La mantisse normalisée peut donc commencer par un 0 tout en conservant la même précision qu'avec 24 bits (53 en double précision).

Ainsi, suivant la notation de Cheney et Kincaid (réf. [5]), les 32 bits de la représentation en simple précision IEEE:

$$(d_1 d_2 d_3 \cdots d_{31} d_{32})_2$$

désignent le nombre décimal:

$$(-1)^{d_1} \times 2^{(d_2 d_3 \cdots d_9)_2} \times 2^{-127} \times (1{,}d_{10} d_{11} \cdots d_{32})_2$$

On remarque immédiatement les différentes composantes: le bit de signe, l'exposant avec un excès de 127 et la mantisse normalisée par l'ajout du 1 manquant.

Remarque 1.6
La norme IEEE traite le nombre réel 0 de façon particulière. □

Exemple 1.10
Les 32 bits suivants (en simple précision IEEE):

$$1100\,0001\,1110\,0000\,0000\,0000\,0000\,0000$$

se décomposent en:

$$\begin{aligned} & (-1)^1 \times 2^{(1000\,0011)_2} \times 2^{-127} \times (1{,}11)_2 \\ = \quad & -2^{131} \times 2^{-127} \times (1 + 2^{-1} + 2^{-2}) \\ = \quad & -16 \times 1{,}75 = -28 \end{aligned}$$

On obtient la représentation du nombre décimal $(30{,}0625)_{10}$ en simple précision au moyen de l'opération inverse. Tout d'abord, la partie entière $(30)_{10}$

devient $(11\,110)_2$ en forme binaire et la partie fractionnaire $(0{,}0625)_{10}$ est tout simplement $(0{,}0001)_2$. Ainsi, on a:

$$(30{,}0625)_{10} = (11\,110{,}0001)_2 = 1{,}111\,000\,01 \times 2^4$$

Dans la dernière expression, la mantisse est normalisée et le bit 1 à la gauche de la virgule n'est pas conservé en mémoire. L'exposant 4 est décalé de 127 pour devenir 131. La représentation de 131 sur 8 bits est $(1000\,0011)_2$. Puisque le nombre $(30{,}0625)_{10}$ est positif, sa représentation en simple précision IEEE sera:

0 1000 0011 1110 0001 0000 0000 0000 000

●●●●●

1.4 Erreurs dues à la représentation

Le fait d'utiliser un nombre limité de bits pour représenter un nombre réel a des conséquences importantes sur la propagation des erreurs. On peut par exemple rechercher quel est le plus petit nombre positif représentable sur 8 bits (voir la figure 1.4). En choisissant la représentation signe et grandeur pour l'exposant, on constate que le nombre doit être positif (0), que son exposant doit être négatif (1), que sa valeur doit être la plus petite possible compte tenu du signe (11) et, enfin, que la mantisse doit être la plus petite possible compte tenu de la normalisation (1000). On obtient:

$$0111\,1000 = \left(1 \times 2^{-1}\right) \times 2^{-3}$$

qui vaut 0,0625. Conséquemment, on ne peut pas représenter tout nombre réel positif plus petit que 0,0625 dans ce système à 8 bits. Un calcul résultant en un nombre plus petit que 0,0625 donnerait lieu à un *sous-dépassement* (*underflow*).

On peut se demander quel serait le nombre suivant. Il s'agit bien sûr de 0111 1001, dont la valeur décimale est 0,070 3125. Il n'est donc pas possible de représenter exactement sur 8 bits les nombres entre 0,0625 et 0,070 3125. Si on cherche par exemple à représenter le nombre réel 0,07, l'ordinateur choisira soit 0111 1000 (c'est-à-dire $(0{,}0625)_{10}$), soit 0111 1001 (c'est-à-dire $(0{,}070\,3125)_{10}$). Le choix dépend de l'utilisation de la *troncature* (*chopping*), qui consiste à choisir systématiquement la borne inférieure, ou de l'*arrondi*

(*rounding*), qui consiste à choisir la borne inférieure seulement si le nombre que l'on souhaite représenter est plus petit que:

$$\frac{0{,}0625 + 0{,}070\,3125}{2}$$

Par ailleurs, le plus grand nombre représentable sur 8 bits est 0011 1111, qui vaut 7,5. Tout calcul résultant en un nombre plus grand que 7,5 donnerait lieu à un débordement.

Cet exemple sur 8 bits exagère grandement les effets de la représentation des nombres réels sur ordinateur. Personne ne songe à faire des calculs numériques avec seulement 8 bits. Si on travaille avec plus de bits, le plus petit nombre représentable devient de plus en plus petit et l'écart entre les nombres représentables diminue également. Les conclusions qui suivent restent cependant vraies.

- Quel que soit le nombre de bits utilisés, il existe un plus petit et un plus grand nombre positifs représentables (de même pour les nombres négatifs). Il existe donc un intervalle fini à l'extérieur duquel on se heurte inévitablement à un débordement ou à un sous-dépassement.
- À l'intérieur de cet intervalle fini, seulement un nombre fini de nombres sont représentables exactement, et on doit recourir à la troncature ou à l'arrondi pour représenter les autres réels.

La représentation en point flottant induit une erreur relative qui dépend du nombre de bits de la mantisse, de l'utilisation de la troncature ou de l'arrondi ainsi que du nombre x que l'on veut représenter. En effet, nous avons vu l'importance sur la précision du nombre de bits de la mantisse et nous avons également constaté qu'un nombre fini seulement de réels peuvent être représentés exactement. L'intervalle entre les nombres représentables varie en longueur selon l'exposant devant la mantisse.

Définition 1.4
La *précision machine* ϵ est la plus grande erreur relative que l'on puisse commettre en représentant un nombre réel sur ordinateur en utilisant la troncature.

Remarque 1.7
La précision machine dépend bien sûr de l'appareil utilisé et du nombre de bits de la mantisse. De plus, si on utilise l'arrondi, la précision machine est tout simplement $\epsilon/2$.□

Théorème 1.1
La précision machine vérifie:

$$\epsilon \leq b^{1-n} \tag{1.13}$$

où b est la base utilisée et n le nombre de bits de la mantisse.

Démonstration:
Soit x, un nombre quelconque. Sa représentation exacte en base b est donc de la forme:

$$x = 0{,}d_1 d_2 d_3 \cdots d_n d_{n+1} d_{n+2} \cdots \times b^l$$

ce qui entraîne que l'erreur absolue commise par troncature est:

$$\begin{aligned}
\Delta x &= 0{,}0000 \cdots 0 d_{n+1} d_{n+2} \cdots \times b^l \\
&= 0{,}d_{n+1} d_{n+2} d_{n+3} \cdots \times b^{l-n} \\
&\leq 0{,}(b-1)(b-1)(b-1) \cdots \times b^{l-n}
\end{aligned}$$

en vertu de l'équation 1.12. L'erreur relative satisfait donc:

$$\begin{aligned}
E_r(x) = \frac{|\Delta x|}{|x|} &\leq \frac{0{,}(b-1)(b-1)(b-1) \cdots \times b^{l-n}}{0{,}d_1 d_2 d_3 \cdots d_n d_{n+1} d_{n+2} \cdots \times b^l} \\
&\leq \frac{0{,}(b-1)(b-1)(b-1) \cdots \times b^{l-n}}{0{,}100\,000 \cdots \times b^l} \\
&\leq 1 \times b^{1-n}
\end{aligned}$$

On note alors que b^{1-n} est une borne supérieure pour la précision machine. Dans ce qui suit, on ne fera plus la distinction entre la précision machine ϵ et sa borne supérieure b^{1-n}, bien que ces deux quantités soient légèrement différentes. □

Il est facile de montrer que cette borne est presque atteinte pour tous les nombres x de développement en base b de la forme:

$$x = 0{,}100\,000 \cdots 0(b-1)(b-1) \cdots \times b^l \tag{1.14}$$

c'est-à-dire dont le 1 est suivi de $(n-1)$ zéros et ensuite d'une infinité de chiffres valant $(b-1)$. L'erreur relative est alors:

$$E_r(x) = \frac{0{,}(b-1)(b-1)\cdots \times b^{l-n}}{0{,}100\,000\cdots 0(b-1)(b-1)\cdots \times b^l}$$

qui est très près de b^{1-n}.

Exemple 1.11
Suivant la norme IEEE, la mantisse d'un réel contient 23 bits en simple précision (52 bits en double précision), mais avec une précision de 24 bits (53 en double précision) puisque, après la normalisation, le premier 1 n'est pas gardé en mémoire. La précision machine vaut alors:

$$2^{1-24} = 0{,}119 \times 10^{-6}$$

en simple précision et

$$2^{1-53} = 0{,}222 \times 10^{-15}$$

en double précision.

● ● ● ● ●

Ce résultat peut être vérifié directement sur ordinateur au moyen de l'algorithme qui suit (voir Chapra et Canale, réf. [4]). Cet algorithme permet de construire une suite de nombres de forme similaire à celle de l'équation 1.14.

Algorithme 1.1: Précision machine

1. $\epsilon = 1$
2. Lorsque $1 + \epsilon > 1$, effectuer $\epsilon = \epsilon/2$
3. $\epsilon = 2 \times \epsilon$ et arrêt
4. Précision machine $= \epsilon$ □

Dans l'algorithme précédent, le nombre $(1 + \epsilon)$ prend successivement les valeurs suivantes:

Forme décimale	Forme binaire
2	$0{,}1 \times 2^2$
1,5	$0{,}11 \times 2^1$
1,25	$0{,}101 \times 2^1$
1,125	$0{,}1001 \times 2^1$
1,0625	$0{,}100\,01 \times 2^1$
$\vdots$	$\vdots$

Cette suite continue jusqu'à ce que le nombre de zéros intercalés dans la représentation binaire soit trop grand et dépasse la longueur de la mantisse. On aura alors $1 + \epsilon = 1$ et l'algorithme s'arrêtera. Il est à noter que la représentation binaire de $1+\epsilon$ est de forme similaire à celle de la relation 1.14. Sur un IBM RISC 6000, la concordance entre le résultat de cet algorithme et l'équation 1.13 est parfaite.

Terminons cette section par deux exemples illustrant les effets parfois étonnants de la représentation binaire.

Exemple 1.12
Si on convertit la fraction décimale 0,1 en sa valeur binaire, on a:

$$\begin{array}{rcll} 0{,}1 \times 2 & = & 0{,}2 & \text{c.-à-d.} \quad d_1 = 0 \\ 0{,}2 \times 2 & = & 0{,}4 & \text{c.-à-d.} \quad d_2 = 0 \\ 0{,}4 \times 2 & = & 0{,}8 & \text{c.-à-d.} \quad d_3 = 0 \\ 0{,}8 \times 2 & = & 1{,}6 & \text{c.-à-d.} \quad d_4 = 1 \\ 0{,}6 \times 2 & = & 1{,}2 & \text{c.-à-d.} \quad d_5 = 1 \\ 0{,}2 \times 2 & = & 0{,}4 & \text{c.-à-d.} \quad d_6 = 0 \\ \vdots & & \vdots & \qquad\qquad \vdots \end{array}$$

ou encore

$$(0{,}1)_{10} = (0{,}000\,110\,011\,00\cdots)_2$$

Ainsi, une fraction ayant un développement décimal fini peut avoir un développement binaire illimité.

● ● ● ● ●

Lorsqu'on utilise un nombre fini de bits dans la mantisse pour représenter $(0,1)_{10}$, l'importance de l'erreur commise dépend du nombre de bits utilisés.

Exemple 1.13
Le problème consiste à sommer 10 000 fois le nombre $(1,0)_{10}$, qui possède un développement binaire fini, et le nombre $(0,1)_{10}$, qui n'a pas de représentation exacte sur un nombre fini de bits. On obtient, sur un IBM RISC 6000, les résultats suivants:

$$10\,000{,}000\,00 \quad \text{et} \quad 999{,}902\,8931$$

en simple précision et:

$$10\,000{,}000\,000\,000\,0000 \quad \text{et} \quad 1\,000{,}000\,014\,901\,161\,19$$

en double précision. Cela démontre l'effet des erreurs de représentation sur ordinateur. Des opérations en apparence identiques donnent, dans un cas, un résultat exact et, dans l'autre, un résultat erroné dont la précision augmente avec le nombre de bits de la mantisse.

● ● ● ● ●

1.5 Arithmétique flottante

Les erreurs ont tendance à se propager et quelques fois à s'amplifier au fil des calculs. Dans cette section, nous suivons l'évolution des erreurs au fil des opérations élémentaires. Afin de simplifier l'exposé, nous utilisons le système décimal, mais les effets décrits valent également pour les autres bases.

Tout nombre réel x s'écrit sous la forme:

$$x = 0{,}d_1d_2d_3 \cdots d_nd_{n+1} \cdots \times 10^l$$

Définition 1.5
Soit x, un nombre réel. On note $\text{fl}(x)$ sa représentation en *notation flottante à n chiffres* définie par:

$$\text{fl}(x) = 0{,}d_1d_2d_3 \cdots d_n \times 10^l$$

Remarque 1.8
La notation flottante d'un nombre dépend du nombre n de chiffres dans la mantisse, mais aussi du procédé retenu pour éliminer les derniers chiffres. Ainsi, la troncature consiste à retrancher les chiffres à partir de la position $n+1$. Avec l'arrondi, on ajoute 5 au $(n+1)^{\text{e}}$ chiffre de la mantisse avant d'effectuer la troncature. La troncature est dite *biaisée*, car on a toujours, pour des nombres positifs, $\text{fl}(x) \leq x$. Par contre, l'arrondi est non biaisé, car on a tour à tour $x \leq \text{fl}(x)$ ou $x \geq \text{fl}(x)$. Nous utilisons l'arrondi dans les exemples qui suivent. □

Remarque 1.9
La norme IEEE recommande l'utilisation de l'arrondi dans la représentation binaire des nombres réels. □

Exemple 1.14
Si on choisit $n = 4$, alors on a:

$$\begin{array}{lcl} \text{fl}(1/3) & = & 0{,}3333 \times 10^0 \\ \text{fl}(\pi) & = & 0{,}3142 \times 10^1 \\ \text{fl}(12{,}4551) & = & 0{,}1246 \times 10^2 \end{array}$$

● ● ● ● ●

1.5.1 Opérations élémentaires

Les opérations élémentaires sont l'addition, la soustraction, la multiplication et la division. Soit x et y, deux nombres réels. On effectue ces opérations en arithmétique flottante de la façon suivante:

$$\begin{array}{lcl} x + y & \rightarrow & \text{fl}(\text{fl}(x) + \text{fl}(y)) \\ x - y & \rightarrow & \text{fl}(\text{fl}(x) - \text{fl}(y)) \\ x \div y & \rightarrow & \text{fl}(\text{fl}(x) \div \text{fl}(y)) \\ x \times y & \rightarrow & \text{fl}(\text{fl}(x) \times \text{fl}(y)) \end{array}$$

En un mot, on doit d'abord représenter les deux opérandes en notation flottante, effectuer l'opération de la façon habituelle et exprimer le résultat en notation flottante.

Exemple 1.15
Si on prend $n = 4$, alors on a:

$$\begin{aligned} 1/3 \times 3 &\rightarrow \text{fl}(\text{fl}(1/3) \times \text{fl}(3)) \\ &= \text{fl}((0{,}3333 \times 10^0) \times (0{,}3000 \times 10^1)) \\ &= \text{fl}(0{,}099\,990\,00 \times 10^1) \\ &= 0{,}9999 \times 10^0 \end{aligned}$$

On remarque une légère perte de précision par rapport à la valeur exacte de cette opération qui est 1.

●●●●●

La multiplication et la division sont particulièrement simples en arithmétique flottante à cause de la loi des exposants.

Exemple 1.16
Toujours avec $n = 4$, effectuer les opérations suivantes:

a) $(0{,}4035 \times 10^6) \times (0{,}1978 \times 10^{-1})$

$$\begin{aligned} &= \text{fl}(0{,}4035 \times 10^6 \times 0{,}1978 \times 10^{-1}) \\ &= \text{fl}(0{,}079\,8123 \times 10^5) \\ &= \text{fl}(0{,}798\,123 \times 10^4) \\ &= 0{,}7981 \times 10^4 \end{aligned}$$

b) $(0{,}567\,89 \times 10^4) \div (0{,}123\,4321 \times 10^{-3})$

$$\begin{aligned} &= \text{fl}(0{,}5679 \times 10^4 \div 0{,}1234 \times 10^{-3}) \\ &= \text{fl}(4{,}602\,106\,969 \times 10^7) \\ &= \text{fl}(0{,}460\,210\,6969 \times 10^8) \\ &= 0{,}4602 \times 10^8 \end{aligned}$$

●●●●●

Par contre, il faut être plus prudent avec l'addition et la soustraction. On ajoute d'abord des zéros à la mantisse du nombre ayant le plus petit exposant de telle sorte que les deux exposants soient égaux. On effectue ensuite l'opération habituelle et on met le résultat en notation flottante.

Exemple 1.17
Toujours avec $n = 4$, effectuer les opérations suivantes:
a) $(0{,}4035 \times 10^6) + (0{,}1978 \times 10^4)$

$$\begin{aligned}
&= \mathrm{fl}(0{,}4035 \times 10^6 + 0{,}1978 \times 10^4)\\
&= \mathrm{fl}(0{,}4035 \times 10^6 + 0{,}001\,978 \times 10^6)\\
&= \mathrm{fl}(0{,}405\,478 \times 10^6)\\
&= 0{,}4055 \times 10^6
\end{aligned}$$

b) $(0{,}567\,89 \times 10^4) - (0{,}123\,4321 \times 10^6)$

$$\begin{aligned}
&= \mathrm{fl}(0{,}5679 \times 10^4 - 0{,}1234 \times 10^6)\\
&= \mathrm{fl}(0{,}005\,679 \times 10^6 - 0{,}1234 \times 10^6)\\
&= -\mathrm{fl}(0{,}117\,72 \times 10^6)\\
&= -0{,}1177 \times 10^6
\end{aligned}$$

On constate qu'il est primordial de décaler la mantisse avant d'effectuer l'addition ou la soustraction.

● ● ● ● ●

Remarque 1.10
Il faut bien remarquer que des opérations mathématiquement équivalentes ne le sont pas forcément en arithmétique flottante. *L'ordre des opérations est très important.* En voici un exemple. □

Exemple 1.18
La propriété de distributivité de la multiplication sur l'addition n'est pas toujours respectée en arithmétique flottante. En effet, en arithmétique exacte:

$$122 \times (333 + 695) = (122 \times 333) + (122 \times 695) = 125\,416$$

En arithmétique flottante avec $n = 3$, on obtient d'une part:

$$\begin{aligned}
122 \times (333 + 695) &= \text{fl}[(0{,}122 \times 10^3) \times \text{fl}(0{,}333 \times 10^3 + 0{,}695 \times 10^3)] \\
&= \text{fl}[(0{,}122 \times 10^3) \times \text{fl}(1{,}028 \times 10^3)] \\
&= \text{fl}[(0{,}122 \times 10^3) \times (0{,}103 \times 10^4)] \\
&= \text{fl}(0{,}012\,566 \times 10^7) \\
&= 0{,}126 \times 10^6
\end{aligned}$$

et d'autre part:

$$\begin{aligned}
(122 \times 333) + (122 \times 695) &= \text{fl}[\text{fl}((0{,}122 \times 10^3) \times (0{,}333 \times 10^3)) \\
&\quad + \text{fl}((0{,}122 \times 10^3) \times (0{,}695 \times 10^3))] \\
&= \text{fl}[\text{fl}(0{,}040\,626 \times 10^6) + \text{fl}(0{,}084\,79 \times 10^6)] \\
&= \text{fl}[(0{,}406 \times 10^5) + (0{,}848 \times 10^5)] \\
&= \text{fl}(1{,}254 \times 10^5) \\
&= 0{,}125 \times 10^6
\end{aligned}$$

On constate donc une légère différence entre les deux résultats, ce qui indique que les deux façons d'effectuer les opérations ne sont pas équivalentes en arithmétique flottante.

◆ ◆ ◆ ◆ ◆

1.5.2 Opérations risquées

Un certain nombre de calculs sont particulièrement sensibles aux erreurs d'arrondi. Nous présentons quelques exemples d'opérations à éviter.

Exemple 1.19

Additionner deux nombres dont les ordres de grandeur sont très différents:

$(0{,}4000 \times 10^4) + (0{,}1000 \times 10^{-2})$

$$\begin{aligned}
&= \text{fl}(0{,}4000 \times 10^4 + 0{,}1000 \times 10^{-2}) \\
&= \text{fl}(0{,}4000 \times 10^4 + 0{,}000\,0001 \times 10^4) \\
&= \text{fl}(0{,}400\,0001 \times 10^4) \\
&= 0{,}4000 \times 10^4
\end{aligned}$$

La loi des exposants et l'arrondi font en sorte que le petit nombre disparaît complètement devant le plus grand. Ce comportement est encore plus en évidence dans l'exemple suivant.

◆ ◆ ◆ ◆ ◆

Exemple 1.20
Calculer une somme de termes positifs:

$$S_n = 1 + \sum_{i=1}^{n} \frac{1}{i^2 + i}$$

On peut évaluer analytiquement cette série en utilisant les fractions partielles. En effet:

$$\begin{aligned} S_n = 1 + \sum_{i=1}^{n} \frac{1}{i^2+i} &= 1 + \sum_{i=1}^{n} \left(\frac{1}{i} - \frac{1}{i+1} \right) \\ &= 1 + (1 - \frac{1}{2}) + (\frac{1}{2} - \frac{1}{3}) + \cdots + (\frac{1}{n} - \frac{1}{n+1}) \\ &= 2 - \frac{1}{n+1} \end{aligned}$$

On peut calculer cette somme tout d'abord directement:

$$S_{1,n} = 1 + \frac{1}{2} + \frac{1}{6} + \cdots + \frac{1}{n^2 + n}$$

et ensuite à rebours:

$$S_{2,n} = \frac{1}{n^2 + n} + \cdots + \frac{1}{6} + \frac{1}{2} + 1$$

Les résultats calculés sur IBM RISC 6000 sont résumés dans le tableau ci-dessous pour différentes valeurs de n.

n	$S_{1,n}$	$S_{2,n}$	Valeur exacte de S_n
99	1,989 999 890	1,990 000 010	1,99
999	1,999 001 026	1,999 000 072	1,999
9999	1,999 791 980	1,999 899 983	1,9999

On remarque que $S_{2,n}$ est plus précis que $S_{1,n}$. Cela est dû au fait que lorsqu'on somme $S_{1,n}$ on accumule dans $S_{1,n}$ les sommes intermédiaires. Conséquemment, $S_{1,n}$ devient de plns en plns grand par rapport à $\frac{1}{i^2+i}$ et on finit

par additionner un très grand et un très petit nombre (ce type d'opération est à éviter). Inversement, lorsqu'on somme $S_{2,n}$, les sommes intermédiaires augmentent, mais les termes qui s'ajoutent au fil de la sommation croissent également et le phénomène précédent ne se produit pas.

◆ ◆ ◆ ◆ ◆

Exemple 1.21
Soustraire deux nombres presque identiques:

$$
\begin{aligned}
(0{,}5678 \times 10^6) - (0{,}5677 \times 10^6) & \\
&= \mathrm{fl}(0{,}5678 \times 10^6 - 0{,}5677 \times 10^6) \\
&= \mathrm{fl}(0{,}0001 \times 10^6) \\
&= 0{,}1000 \times 10^3
\end{aligned}
$$

La soustraction de ces 2 nombres de valeur très proche fait apparaître trois 0 non significatifs dans la mantisse du résultat. On appelle ce phénomène l'*élimination par soustraction des chiffres significatifs*. L'exemple suivant en illustre les conséquences possibles.

◆ ◆ ◆ ◆ ◆

Exemple 1.22 (Chapra et Canale, réf. [4])
Calculer les racines de:

$$ax^2 + bx + c$$

qui sont bien sûr:

$$r_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{et} \quad r_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

On considère le cas où $a = 1$, $b = 3000{,}001$ et $c = 3$. Les racines exactes sont $r_1 = -0{,}001$ et $r_2 = -3000$. Un calcul en simple précision (norme IEEE) donne $r_1 = -0{,}000\,988\,281\,33$ et $r_2 = -3000{,}0000$, dont l'erreur relative est respectivement de 1,17 % et de 0,0 %. Le calcul de r_2 ne pose aucune difficulté particulière. L'erreur relative liée à r_1 provient de l'addition de $(-b)$, qui vaut $-3000{,}000\,977$, et de $\sqrt{b^2 - 4ac}$, qui vaut 2999,999 023. Cette

opération revient à soustraire des nombres très voisins. Pour éviter cela, on peut multiplier r_1 par son conjugné et calculer:

$$r_1 = \left(\frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)\left(\frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}}\right) = \frac{-2c}{b + \sqrt{b^2 - 4ac}}$$

On obtient de cette manière:

$$r_1 = -0{,}001\,000\,000\,047$$

et une erreur relative extrêmement faible.

●●●●●

Exemple 1.23

Évaluer une série de nombres de signes alternés (Chapra et Canale, réf [4]).

Nous verrons un peu plus loin que le développement de Taylor de la fonction e^{-x} autour de 0 est:

$$\begin{aligned} e^{-x} &= 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots + \frac{(-1)^n x^n}{n!} + \frac{(-1)^{n+1} x^{n+1}}{(n+1)!} + \cdots \\ &= S_n + \frac{(-1)^{n+1} x^{n+1}}{(n+1)!} + \cdots \end{aligned} \qquad (1.15)$$

On remarque que les termes de la série changent constamment de signe. Dans l'expression 1.15, S_n désigne la somme des $(n + 1)$ premiers termes de la série. Le tableau ci-dessous présente quelques résultats intermédiaires obtenus lors de l'évaluation de e^{-10}, qui consiste à poser tout simplement $x = 10$ dans l'équation 1.15.

n	$x^n/n!$	Somme partielle S_n
0	+1,000 000 00	+1,000 000 000
1	−10,000 0000	−9,000 000 000
2	+50,000 0000	+41,000 000 00
3	−166,666 718	−125,666 6718
4	+416,666 687	+291,000 0000
5	−833,333 374	−542,333 3740
10	+2755,732 17	+1342,587 402
20	+41,103 1875	+13,396 751 40
30	$+0{,}376\,998\,9125 \times 10^{-2}$	$+0{,}852\,284\,9530 \times 10^{-3}$
31	$-0{,}121\,612\,5558 \times 10^{-2}$	$-0{,}363\,840\,6051 \times 10^{-3}$
32	$+0{,}380\,039\,2442 \times 10^{-3}$	$+0{,}161\,986\,3906 \times 10^{-4}$
33	$-0{,}115\,163\,4060 \times 10^{-3}$	$-0{,}989\,647\,6690 \times 10^{-4}$
34	$+0{,}338\,715\,9086 \times 10^{-4}$	$-0{,}650\,931\,7609 \times 10^{-4}$
35	$-0{,}967\,759\,7518 \times 10^{-5}$	$-0{,}747\,707\,7452 \times 10^{-4}$
40	$+0{,}122\,561\,8007 \times 10^{-7}$	$-0{,}726\,567\,8660 \times 10^{-4}$
41	$-0{,}298\,931\,2131 \times 10^{-8}$	$-0{,}726\,576\,6908 \times 10^{-4}$
42	$+0{,}711\,740\,9990 \times 10^{-9}$	$-0{,}726\,569\,5604 \times 10^{-4}$
43	$-0{,}165\,521\,1662 \times 10^{-9}$	$-0{,}726\,571\,2338 \times 10^{-4}$
44	$+0{,}376\,184\,4655 \times 10^{-10}$	$-0{,}726\,570\,8700 \times 10^{-4}$
45	$-0{,}835\,965\,4982 \times 10^{-11}$	$-0{,}726\,570\,9428 \times 10^{-4}$

Voilà un exemple parfait de calcul instable. On remarque immédiatement que le résultat final est faux, car on obtient une valeur négative pour l'exponentielle dont la valeur exacte est $0{,}453\,9993 \times 10^{-4}$. On voit aussi qu'il est inutile de continuer pour des valeurs de n plus grandes que 45, car la valeur de $x^n/n!$ est trop petite pour modifier substantiellement le résultat. Cela revient à additionner un très grand et un très petit nombre. On constate enfin que le phénomène d'élimination par soustraction des chiffres significatifs se produit de façon systématique en raison de l'alternance des signes.

On peut contourner ces difficultés en évaluant la série de manière différente. Une possibilité consiste à utiliser l'algorithme de Horner pour l'évaluation des polynômes.

• • • • •

1.5.3 Évaluation des polynômes

Il est très fréquent d'avoir à évaluer des polynômes de degré élevé en analyse numérique. Il est donc important de pouvoir les évaluer rapidement et de la façon la plus stable possible du point de vue de l'arithmétique flottante. C'est ce que permet l'*algorithme de Horner* appelé aussi *algorithme de multiplication imbriquée*. Pour évaluer un polynôme de la forme:

$$p(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n$$

en un point x quelconque, il suffit de regrouper judicieusement les termes de la façon suivante:

$$p(x) = a_0 + x(a_1 + x(a_2 + x(a_3 + \cdots + x(a_{n-1} + a_nx)\cdots))) \qquad (1.16)$$

Voyons l'évaluation d'un polynôme à l'exemple suivant.

Exemple 1.24
Soit le polynôme:

$$p(x) = 2 + 4x + 5x^2 + 3x^3$$

qui nécessite 6 multiplications et 3 additions. En suivant le mode de regroupement de l'équation 1.16, on obtient:

$$p(x) = 2 + x(4 + x(5 + 3x))$$

qui nécessite seulement 3 multiplications et 3 additions (et aucune élévation de puissance). On réduit donc substantiellement le nombre d'opérations nécessaires et, de plus, cette nouvelle expression est moins sensible aux effets de l'arithmétique flottante.

● ● ● ● ●

La méthode de Horner est facile à programmer grâce à l'algorithme suivant.

Algorithme 1.2: Méthode de Horner

1. Étant donné les coefficients a_i d'un polynôme de degré n
2. Étant donné une abscisse x

3. Étant donné la variable p_n qui contiendra la valeur du polynôme en x
4. $p_n = a_n$
5. Pour $i = n, n-1, n-2, \cdots, 2, 1$:

$$p_n = a_{i-1} + p_n x$$

6. Écrire x, $p(x) = p_n$ □

Exemple 1.25
La méthode de Horner peut servir à reprendre le calcul e^{-10} par la série alternée 1.15. Les coefficients du polynôme sont:

$$a_i = \frac{(-1)^i}{i!}$$

Pour le polynôme de degré $n = 45$, on obtient la valeur approximative $0{,}453\,999 \times 10^{-4}$, qui est très près de la valeur exacte et qui ne change plus pour les valeurs suivantes de n.

●●●●●

1.6 Erreurs de troncature

Les erreurs de troncature constituent la principale catégorie d'erreurs. Tout au long de ce manuel, nous abordons des méthodes de résolution qui comportent des erreurs de troncature plus ou moins importantes. L'*ordre* d'une méthode dépend du nombre de termes utilisés dans les développements de Taylor appropriés. Il est donc essentiel de revoir en détail le développement de Taylor, car il constitue l'outil fondamental de l'analyse numérique.

Remarque 1.11
Il est important de ne pas confondre les erreurs de troncature traitées dans cette section et la troncature utilisée pour la représentation des nombres sur ordinateur. □

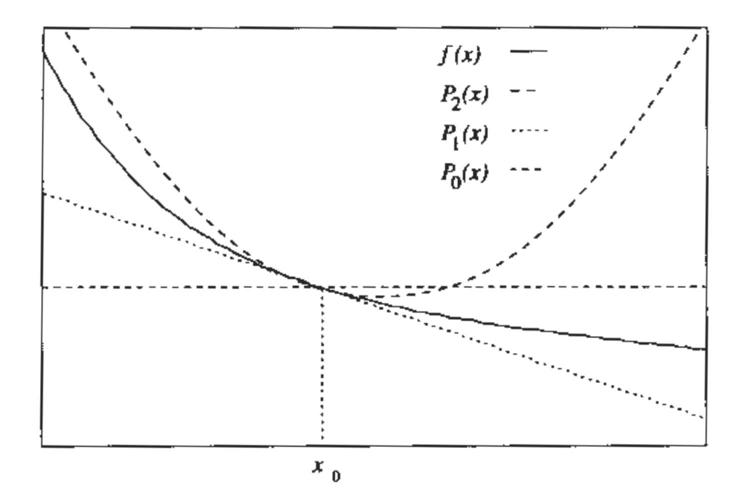


Figure 1.5: Approximation de $f(x)$ au voisinage de x_0

1.6.1 Développement de Taylor en une variable

Il existe plusieurs façons d'introduire le développement de Taylor. Une façon très simple consiste à le présenter formellement comme un problème d'approximation au voisinage d'un point x_0. On se demande alors quel est le polynôme de degré 0 (noté $P_0(x)$) qui donne la meilleure approximation d'une fonction $f(x)$ donnée dans le voisinage du point x_0.

Selon la figure 1.5, ce polynôme est:

$$P_0(x) = f(x_0)$$

On peut pousser plus loin l'analyse et chercher le meilleur polynôme de degré 1 de la forme:

$$P_1(x) = a_0 + a_1(x - x_0) \tag{1.17}$$

On pourrait tout aussi bien chercher un polynôme de forme plus classique:

$$P_1(x) = b_0 + b_1\, x$$

Ces deux expressions sont équivalentes et aboutissent au même résultat. La forme 1.17 est plus pratique et plus naturelle puisqu'elle s'articule autour du point x_0.

On doit introduire deux conditions pour déterminer les deux constantes. Intuitivement, la meilleure droite (polynôme de degré 1) est celle qui passe

par $(x_0, f(x_0))$ et dont la pente est celle de la fonction $f(x)$ en x_0, ce qui entraîne que:

$$P_1(x_0) = f(x_0) \tag{1.18}$$
$$P_1'(x_0) = f'(x_0) \tag{1.19}$$

ou encore

$$\begin{aligned} a_0 &= f(x_0) \\ a_1 &= f'(x_0) \end{aligned}$$

Il en résulte l'expression suivante:

$$P_1(x) = f(x_0) + f'(x_0)\,(x - x_0)$$

On peut bien sûr poursuivre ce raisonnement à condition que la fonction $f(x)$ soit suffisamment dérivable. Dans le cas d'un polynôme de degré 2, on imposerait:

$$P_2(x_0) = f(x_0) \tag{1.20}$$
$$P_2'(x_0) = f'(x_0) \tag{1.21}$$
$$P_2''(x_0) = f''(x_0) \tag{1.22}$$

pour obtenir facilement:

$$P_2(x) = f(x_0) + f'(x_0)\,(x - x_0) + \frac{f''(x_0)\,(x - x_0)^2}{2}$$

En poursuivant ce raisonnement jusqu'à l'ordre n, c'est-à-dire en imposant l'égalité des dérivées jusqu'à l'ordre n, on obtient le ***polynôme de Taylor de degré n*** de la fonction $f(x)$ autour de x_0.

Définition 1.6
Le polynôme de Taylor de degré n de la fonction $f(x)$ autour de x_0 est défini par:

$$\begin{aligned} P_n(x) &= f(x_0) + f'(x_0)\,(x - x_0) + \frac{f''(x_0)\,(x - x_0)^2}{2!} \\ &+ \frac{f'''(x_0)\,(x - x_0)^3}{3!} + \cdots + \frac{f^{(n)}(x_0)\,(x - x_0)^n}{n!} \end{aligned} \tag{1.23}$$

où $f^{(n)}(x_0)$ désigne la dérivée d'ordre n de $f(x)$ en x_0.

Ce polynôme donne une approximation de la fonction $f(x)$ au voisinage de x_0. Il n'y a cependant pas égalité en ce sens que si on utilise l'équation 1.23 pour estimer $f(x)$ on commettra une erreur.

Remarque 1.12
On choisit généralement le point x_0 où on développe le polynôme de Taylor de façon à ce que l'on puisse facilement évaluer la fonction $f(x)$ ainsi que ses dérivées. □

Le résultat suivant quantifie l'erreur commise lorsqu'on utilise le polynôme de Taylor (Thomas et Finney, réf. [22]).

Théorème 1.2
Soit $f(x)$, une fonction dont les dérivées jusqu'à l'ordre $(n+1)$ existent au voisinage du point x_0. On a l'égalité suivante:

$$f(x) = P_n(x) + R_n(x) \tag{1.24}$$

où $P_n(x)$ est le polynôme de Taylor 1.23 et $R_n(x)$ est l'erreur commise, qui est donnée par:

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))\,(x - x_0)^{(n+1)}}{(n+1)!} \tag{1.25}$$

pour un certain $\xi(x)$ compris entre x_0 et x. □

Remarque 1.13

1. L'équation 1.24 est une *égalité* et ne devient une approximation que lorsque le terme d'erreur est négligé.
2. Le terme d'erreur de l'équation 1.25 devient de plus en plus grand lorsque x s'éloigne de x_0 en vertu du terme $(x - x_0)^{(n+1)}$ (voir la figure 1.5).
3. Inversement, pour une valeur de x près de x_0, le terme d'erreur de l'équation 1.25 est de plus en plus petit lorsque n augmente.
4. On sait que le point $\xi(x)$ existe et qu'il varie avec x, mais on ne connaît pas sa valeur exacte. Il n'est donc pas possible d'évaluer le terme d'erreur exactement. On peut tout au plus lui trouver une borne supérieure dans la plupart des cas.

5. On commet une *erreur de troncature* chaque fois que l'on utilise le développement de Taylor et que l'on néglige le terme d'erreur de l'équation 1.25. □

Un cas particulier important du théorème précédent est le premier *théorème de la moyenne*, qui équivaut à poser $n = 0$ dans le développement de Taylor.

Corollaire 1.1
Soit $f(x)$, une fonction dérivable dans l'intervalle $[x_0, x]$. Alors il existe ξ dans $[x_0, x]$ tel que:

$$f(x) = f(x_0) + f'(\xi)(x - x_0)$$

qui s'écrit également sous la forme:

$$f(x) - f(x_0) = f'(\xi)(x - x_0) \quad \square \tag{1.26}$$

On crée une forme plus pratique du développement de Taylor en remplaçant x par $x_0 + h$ ou encore l'expression $x - x_0$ par h. On obtient ainsi:

$$f(x_0 + h) = P_n(h) + R_n(h) \tag{1.27}$$

où

$$\begin{aligned} P_n(h) &= f(x_0) + f'(x_0)\, h + \frac{f''(x_0)\, h^2}{2!} \\ &+ \frac{f'''(x_0)\, h^3}{3!} + \cdots + \frac{f^{(n)}(x_0)\, h^n}{n!} \end{aligned} \tag{1.28}$$

et

$$R_n(h) = \frac{f^{(n+1)}(\xi(h))\, h^{(n+1)}}{(n+1)!} \tag{1.29}$$

pour $\xi(h)$ compris entre x_0 et $x_0 + h$.

Exemple 1.26
On considère la fonction $f(x) = e^x$ au voisinage de $x_0 = 0$. Puisque toutes les dérivées de e^x sont égales à e^x et valent 1 en $x_0 = 0$, le développement

de Taylor de degré n devient:

$$e^{x_0+h} = e^h \simeq P_n(h) = 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + \cdots + \frac{h^n}{n!}$$

et l'expression du terme d'errenr de l'équation 1.29 est:

$$R_n(h) = \frac{e^{\xi(h)}h^{n+1}}{(n+1)!}$$

où $\xi(h) \in [0, h]$. On peut même déterminer une borne supérieure pour le terme d'errenr. La fonction exponentielle étant croissante, on a:

$$e^{\xi(h)} \leq e^h$$

dans l'intervalle $[0, h]$ et on conclut que:

$$R_n(h) \leq \frac{e^h h^{n+1}}{(n+1)!} \tag{1.30}$$

On peut utiliser ce développement pour estimer la valeur de $e^{0,1}$ en prenant $h = 0,1$.

n	$P_n(0,1)$	Erreur absolue	Nombre de chiffres significatifs	Borne pour le terme d'erreur
0	1,000 0000	$0,105 \times 10^0$	1	$0,111 \times 10^0$
1	1,100 0000	$0,517 \times 10^{-2}$	2	$0,552 \times 10^{-2}$
2	1,105 0000	$0,171 \times 10^{-3}$	4	$0,184 \times 10^{-3}$
3	1,105 1667	$0,420 \times 10^{-5}$	6	$0,460 \times 10^{-5}$

On obtient l'erreur absolue simplement en comparant le résultat avec la valeur exacte 1,105 170 918, tandis que la borne supérieure de l'erreur provient de l'équation 1.30 avec $h = 0,1$. Enfin, si on prend $h = 0,05$ et $n = 3$ pour estimer $e^{0,05}$, on obtient:

$$P_3(0,05) = 1,051\,270\,833$$

et une erreur absolue d'environ $0,263 \times 10^{-6}$. On remarque de plus que le rapport des erreurs absolues liées à $P_3(h)$ est:

$$\frac{|P_3(0,1) - e^{0,1}|}{|P_3(0,05) - e^{0,05}|} = \frac{0,4245 \times 10^{-5}}{0,263 \times 10^{-6}} = 16,14$$

La valeur de ce rapport n'est pas fortuite. La définition suivante permet de comprendre d'où provient cette valeur (Bourdeau et Gélinas, réf. [1]).

•••••

Définition 1.7
Une fonction $f(h)$ est un ***grand ordre*** de h^n au voisinage de 0 (noté $f(x) = O(h^n)$) s'il existe une constante positive C telle que:

$$\left|\frac{f(h)}{h^n}\right| \leq C$$

au voisinage de 0.

Bien qu'imprécise, cette définition exprime assez bien les caractéristiques d'une fonction de type $O(h^n)$. Lorsque h est assez petit, la fonction $O(h^n)$ décroît comme Ch^n. Plus n est grand, plus la décroissance est rapide. Ainsi, une fonction $O(h^3)$ décroît plus vite qu'une fonction $O(h^2)$, qui elle-même décroît plus vite qu'une fonction $O(h)$. Pour avoir une idée du comportement d'une fonction de type $O(h^n)$, il suffit de remarquer que, lorsque h est divisé par 2, la fonction $O(h^n)$ diminue selon un facteur approximatif de 2^n. En effet, si on remplace h par $h/2$ dans Ch^n, on obtient:

$$C\ (\frac{h}{2})^n = \frac{Ch^n}{2^n}$$

Remarque 1.14
Le terme d'erreur du polynôme de Taylor de degré n est généralement de type $O(h^{n+1})$. Cela explique le rapport de 16,14 obtenu dans l'exemple précédent. En effet, on y trouve un polynôme de Taylor de degré 3 dont le terme d'erreur est de type $O(h^4)$. En passant de $h = 0{,}1$ à $h = 0{,}05$, on divise h par un facteur de 2, d'où une diminution selon un facteur de $2^4 = 16$ de l'errenr. Bien sûr, le facteur de 16 est approximatif et n'est atteint qu'à des valeurs de h très petites. Dans le cas général, on note:

$$f(x_0 + h) = P_n(h) + O(h^{n+1}) \quad \square$$

Définition 1.8
Une approximation dont le terme d'erreur est un grand ordre de h^n ($O(h^n)$) est dite *d'ordre n*.

Remarque 1.15
Suivant cette définition, le polynôme de Taylor de degré n est généralement (*mais pas toujours*) une approximation d'ordre $(n+1)$ de $f(x)$. Par exemple, le développement de Taylor de degré n de e^x autour de $x = 0$ est d'ordre $(n+1)$. □

Remarque 1.16
Dans certains manuels, il y a confusion entre le degré et l'ordre du polynôme de Taylor. Il faut s'assurer de bien distinguer ces deux notions. □

Exemple 1.27
Calculer le développement de Taylor d'ordre 5 de la fonction $\sin x$ autour de $x_0 = 0$. Les dérivées de la fonction sont respectivement:

$$\begin{array}{rclrcl} f(x) & = & \sin x, & f(0) & = & 0 \\ f'(x) & = & \cos x, & f'(0) & = & 1 \\ f''(x) & = & -\sin x, & f''(0) & = & 0 \\ f'''(x) & = & -\cos x, & f'''(0) & = & -1 \\ f''''(x) & = & \sin x, & f''''(0) & = & 0 \\ f'''''(x) & = & \cos x & & & \end{array}$$

Le développement de Taylor est donc:

$$\sin(x_0 + h) = \sin(h) = h - \frac{h^3}{3!} + \frac{\cos(\xi(h))h^5}{5!}$$

Il suffit de calculer le polynôme de Taylor de degré 3 ($P_3(h)$) pour obtenir une approximation d'ordre 5 de la fonction $\sin h$. Puisque la fonction $\cos x$ est bornée en valeur absolue par 1, on note immédiatement que:

$$|R_3(h)| \leq \frac{h^5}{5!}$$

Si on prend maintenant $h = 0{,}1$, on peut obtenir l'approximation:

$$\sin(0{,}1) \simeq 0{,}1 - \frac{(0{,}1)^3}{3!} = 0{,}099\,833\,333$$

soit une erreur absolue de $0{,}8332 \times 10^{-7}$. Il est à noter que la borne supérieure de l'erreur vaut $0{,}8333 \times 10^{-7}$, ce qui est très près de la valeur exacte. Si on prend $h = 0{,}2$, on trouve:

$$\sin(0{,}2) \simeq 0{,}2 - \frac{(0{,}2)^3}{3!} = 0{,}198\,666\,6667$$

et une erreur absolue de $0{,}2664 \times 10^{-5}$. On remarque de plus que le rapport entre les deux erreurs absolues est:

$$\frac{0{,}2664 \times 10^{-5}}{0{,}8332 \times 10^{-7}} = 31{,}97$$

ce qui confirme que cette approximation est bien d'ordre 5. En prenant une valeur de h deux fois plus grande, on trouve une erreur à peu près $2^5 = 32$ fois plus grande. Cet exemple montre que le polynôme de Taylor de degré 3 de la fonction $f(x) = \sin x$ est d'ordre 5.

● ● ● ● ●

1.6.2 Développement de Taylor en plusieurs variables

On peut reprendre, dans le cas de plusieurs variables, le raisonnement qui a mené au développement de Taylor d'une variable. Nous nous limitons, pour les fins de l'exposé, à trois variables, le cas général étant similaire.

Théorème 1.3
Soit $f(x_1, x_2, x_3)$, une fonction de trois variables, que l'on suppose suffisamment différentiable. On a alors:

$$f(x_1 + h_1, x_2 + h_2, x_3 + h_3) = f(x_1, x_2, x_3)$$

$$+\left(\frac{\partial f(x_1, x_2, x_3)}{\partial x_1}h_1 + \frac{\partial f(x_1, x_2, x_3)}{\partial x_2}h_2 + \frac{\partial f(x_1, x_2, x_3)}{\partial x_3}h_3\right)$$

$$+\frac{1}{2!}\left(\frac{\partial^2 f(x_1, x_2, x_3)}{\partial x_1^2}h_1^2 + \frac{\partial^2 f(x_1, x_2, x_3)}{\partial x_2^2}h_2^2 + \frac{\partial^2 f(x_1, x_2, x_3)}{\partial x_3^2}h_3^2\right)$$

$$+\left(\frac{\partial^2 f(x_1, x_2, x_3)}{\partial x_1 \partial x_2}h_1 h_2 + \frac{\partial^2 f(x_1, x_2, x_3)}{\partial x_1 \partial x_3}h_1 h_3 + \frac{\partial^2 f(x_1, x_2, x_3)}{\partial x_2 \partial x_3}h_2 h_3\right)$$

$$+\cdots$$

Les termes suivants (désignés par les pointillés) font intervenir les différentes dérivées partielles d'ordre 3, 4, 5 $\cdots$ de la fonction $f(x_1, x_2, x_3)$. On voit bien la similitude avec le cas d'une variable. En pratique, on utilise principalement le développement de degré 1, qui ne fait intervenir que les dérivées partielles d'ordre 1. □

Exemple 1.28
Soit la fonction de deux variables:

$$f(x_1, x_2) = x_1^2 + x_1 \sin x_2$$

que l'on développe autour de $(x_1, x_2) = (1\,,\,0)$. On a alors $f(1\,,\,0) = 1$. De plus, les dérivées partielles du premier ordre de f sont:

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1 + \sin x_2 \qquad \frac{\partial f(x_1, x_2)}{\partial x_2} = x_1 \cos x_2$$

et celles du deuxième ordre sont:

$$\frac{\partial^2 f(x_1, x_2)}{\partial x_1^2} = 2 \qquad \frac{\partial^2 f(x_1, x_2)}{\partial x_2^2} = -x_1 \sin x_2 \qquad \frac{\partial^2 f(x_1, x_2)}{\partial x_1 \partial x_2} = \cos x_2$$

Au point $(1\,,0)$, ces dérivées partielles valent:

$$\frac{\partial f(1\,,0)}{\partial x_1} = 2 \qquad \frac{\partial f(1\,,0)}{\partial x_2} = 1$$

et

$$\frac{\partial^2 f(1\,,0)}{\partial x_1^2} = 2 \qquad \frac{\partial^2 f(1\,,0)}{\partial x_2^2} = 0 \qquad \frac{\partial^2 f(1\,,0)}{\partial x_1 \partial x_2} = 1$$

Le développement de Taylor de degré 2 de cette fonction de deux variables autour du point $(1\,,0)$ est donc:

$$f(1 + h_1, 0 + h_2) \simeq 1 + 2h_1 + 1h_2 + \frac{1}{2}(2h_1^2 + 0h_2^2) + (1h_1h_2)$$

c'est-à-dire

$$f(1 + h_1, h_2) \simeq 1 + 2h_1 + h_2 + h_1^2 + h_1h_2$$

En choisissant par exemple $h_1 = h_2 = 0{,}1$, on obtient l'approximation snivante:

$$f(1{,}1\,,0{,}1) \simeq 1{,}32$$

qui est proche de la valenr exacte 1,319 816 758 avec une erreur absolue d'environ 0,000 183. Si on prend maintenant $h_1 = h_2 = 0{,}05$, on obtient une approximation de $f(1{,}05\,,0{,}05)$ qui vaut 1,155. L'erreur absolue dans ce dernier cas est d'environ 0,000 021 825, qui est environ 8,4 fois plus petite qu'avec $h_1 = h_2 = 0{,}1$. Ce facteur de 8 s'explique par le choix d'un développement de degré 2 (et d'ordre 3) et par la division des valeurs de h_1 et h_2 par 2.

● ● ● ● ●

1.6.3 Propagation d'erreurs dans le cas général

Nous approfondissons dans cette section plusieurs notions vues plus haut. Que peut-on dire, par exemple, de la précision des résultats obtenus lorsqu'on additionne ou qu'on multiplie des valeurs connues avec une précision limitée? Plus généralement, si on a:

$$\begin{aligned} x &= x^* \pm \Delta x \\ y &= y^* \pm \Delta y \end{aligned}$$

quelle sera la précision de la fonction d'une variable $f(x^*)$ ou de la fonction de deux variables $g(x^*, y^*)$? Ici encore, le développement de Taylor apporte

une solution. Considérons d'abord le cas d'une variable. Une quantité x inconnue est approchée par une valeur approximative x^* avec une erreur absolue Δx. On estime la valeur inconnue $f(x)$ par l'approximation $f(x^*)$. L'erreur absolue liée à ce résultat est:

$$\Delta f = |f(x) - f(x^*)|$$

On a de plus:

$$f(x) = f(x^* \pm \Delta x) = f(x^*) \pm f'(x^*)\, \Delta x + O((\Delta x)^2)$$

En négligeant les termes d'ordre plus grand ou égal à 2, on obtient:

$$\Delta f \simeq |f'(x^*)|\, \Delta x$$

que l'on peut également écrire:

$$f(x) = f(x^*) \pm |f'(x^*)|\Delta x \tag{1.31}$$

Exemple 1.29
On a mesuré la longueur d'un côté d'une boîte cubique et obtenu $l^* = 10{,}2$ cm avec une précision de l'ordre du millimètre ($\Delta l = 0{,}1$ cm). On cherche le volume v de cette boîte. Dans ce cas, $f(l) = l^3 = v$ et l'erreur liée au volume est:

$$\Delta v = |f'(l^*)|\, \Delta l = 3\,(10{,}2)^2 \times 0{,}1 = 31{,}212 \leq 0{,}5 \times 10^2$$

La valeur approximative du volume est $(10{,}2)^3 = 1061{,}2$ cm^3, dont seuls les deux premiers chiffres sont significatifs.

●●●●●

On traite les fonctions de plusieurs variables en faisant appel au développement de Taylor en plusieurs variables. Nous donnons le résultat en dimension 3 seulement, car le cas général ne pose aucune difficulté supplémentaire.

Théorème 1.4
Soit $f(x, y, z)$, une fonction de trois variables x, y et z dont on estime la valeur par x^*, y^* et z^* avec une précision de Δx, de Δy et de Δz respectivement. L'erreur absolue Δf est donnée par:

$$\begin{aligned} \Delta f &= \left|\frac{\partial f(x^*, y^*, z^*)}{\partial x}\right| \Delta x + \left|\frac{\partial f(x^*, y^*, z^*)}{\partial y}\right| \Delta y \\ &+ \left|\frac{\partial f(x^*, y^*, z^*)}{\partial z}\right| \Delta z \qquad \square \end{aligned} \tag{1.32}$$

Exemple 1.30
Un signal électrique est donné par:

$$V = A \sin(\omega t - \phi)$$

où V est la tension, A est l'amplitude du signal ($A^* = 100$ V), ω est la fréquence ($\omega^* = 3$ rad/s), ϕ est le déphasage ($\phi^* = 0{,}55$ rad) et t est le temps ($t^* = 0{,}001$ s). En supposant que A et ω sont connus exactement ($A = A^*$, $\omega = \omega^*$) et que ϕ^* et t^* possèdent respectivement 2 et 1 chiffres significatifs, il s'agit d'évaluer l'erreur absolue liée à V ainsi que le nombre de chiffres significatifs.

Puisque A et ω sont connns exactement, on sait immédiatement que ΔA et $\Delta\omega$ sont nuls et qu'ils n'ont aucune contribution à l'erreur liée à V. Par ailleurs:

$$\begin{aligned} \Delta t &= 0{,}5 \times 10^{-3} \\ \Delta \phi &= 0{,}5 \times 10^{-2} \end{aligned}$$

et l'erreur totale est:

$$\Delta V = \left|\frac{\partial V(t^*, \phi^*)}{\partial t}\right| \Delta t + \left|\frac{\partial V(t^*, \phi^*)}{\partial \phi}\right| \Delta \phi$$

c'est-à-dire

$$\Delta V = |A^*\omega^* \cos(\omega^* t^* - \phi^*)| \times (0{,}5 \times 10^{-3}) + |-A^* \cos(\omega^* t^* - \phi^*)| \times (0{,}5 \times 10^{-2})$$

ce qui donne:

$$\begin{aligned} \Delta V &= 256{,}226\,662\,35 \times (0{,}5 \times 10^{-3}) + |-85{,}408\,874\,5| \times (0{,}5 \times 10^{-2}) \\ &= = 0{,}555\,157\,684 \end{aligned}$$

La tension approximative V^* est bien sûr:

$$V^* = A^* \sin(\omega^* t^* - \phi^*) = -52{,}012\,730\,71$$

et puisque $\Delta V \leq 0{,}5 \times 10^1$, ce nombre n'a qu'un seul chiffre significatif.

●●●●●

Quelques cas particuliers méritent de l'attention. De l'équation 1.32, on peut déduire la façon dont se propagent les erreurs dans les opérations élémentaires. En effet, en prenant par exemple $f(x, y) = x/y$, on trouve:

$$\Delta f = \left|\frac{1}{y}\right| \Delta x + \left|\frac{-x}{y^2}\right| \Delta y = \frac{|y|\Delta x + |x|\Delta y}{y^2}$$

ou encore

$$\Delta(x \div y) = \frac{|y|\Delta x + |x|\Delta y}{y^2}$$

On obtient ainsi le tableau suivant à partir de l'équation 1.32.

$\Delta(x+y)$	$=$	$\Delta x + \Delta y$
$\Delta(x-y)$	$=$	$\Delta x + \Delta y$
$\Delta(x \times y)$	$=$	$\|y\|\Delta x + \|x\|\Delta y$
$\Delta(x \div y)$	$=$	$\frac{\|y\|\Delta x + \|x\|\Delta y}{y^2}$

(1.33)

On remarque que les erreurs absolues pour la soustraction s'*additionnent*. Le tableau montre également la similitude entre la propagation d'erreurs et la différentiation d'une somme, d'une différence, d'un produit et d'un quotient de deux fonctions.

1.7 Exercices

1. Tous les chiffres des nombres suivants sont significatifs. Donner une borne supérieure de l'erreur absolue et estimer l'erreur relative.

a) 0,1234	b) 8,760	c) 3,141 56
d) $0{,}112\,35 \times 10^{-3}$	e) 8,000	f) $0{,}223\,56 \times 10^{8}$

2. Exprimer les nombres décimaux suivants en représentation binaire classique.

   a) 32 b) 125 c) 1231 d) 876 e) 999 f) 12 345

3. Exprimer les entiers signés 125, -125, 0, 175 et -100 dans une forme binaire sur 8 bits en utilisant:

   a) la représentation signe et grandeur.

   b) le complément à 2.

   c) la représentation par excès ($d = 2^7$).

4. Traduire les nombres binaires 0000 0011, 1000 0001 et 1111 1111 dans la forme décimale selon que la représentation utilisée est:

   a) la représentation binaire classique.

   b) la représentation signe et grandeur.

   c) le complément à 2.

   d) la représentation par excès ($d = 2^7$).

5. Pour représenter les nombres réels, considérer un mot de 16 bits dont 1 exprime le signe du nombre, 5, l'exposant et 10, la mantisse.

   a) Représenter le plus petit et le plus grand nombre positifs en utilisant le complément à 2 pour l'exposant.

   b) Déterminer la précision machine.

6. Convertir en forme binaire les fractions décimales suivantes.

   a) 0,5 b) 0,2 c) 0,9 d) 1/3 e) 0,25 f) 3/8

7. Un ordinateur fictif représente les nombres réels sur 32 bits dans l'ordre suivant:

   1 bit pour le signe du nombre (0 = positif, 1 = négatif);

   7 bits ponr l'exposant en représentation signe et grandeur;

   24 bits pour la mantisse (normalisée).

   a) Que représentent (en base 10) les 32 bits:

   $$1000\,1011\,1010\,1000\,0010\,0000\,0000\,0000$$

   b) Donner nne borne supérieure de l'erreur relative liée à cette représentation (on suppose que l'ordinateur utilise la troncature).

   c) Donner l'expression binaire sur 32 bits du plus petit nombre réel positif représentable et donner sa valeur en base 10.

8. Convertir les nombres suivants en simple précision IEEE.

   a) $-52{,}234\,375$ b) $7112{,}0$ c) $16{,}2$

   Vérifier les réponses en les retransformant en nombres décimaux. Évaluer l'erreur de représentation commise en c).

9. Donner la représentation en notation flottante en base 10 des nombres suivants (arrondir en conservant 4 chiffres dans la mantisse).

a) e	b) $1/6$	c) $2/3$
d) $12{,}487 \times 10^5$	e) $213\,456$	f) $2000{,}1$

10. Montrer que la loi d'associativité de l'addition n'est pas toujours respectée en arithmétique flottante. Utiliser l'arithmétique flottante à 3 chiffres et les nombres suivants: $x = 0{,}854 \times 10^3$, $y = 0{,}251 \times 10^3$ et $z = 0{,}852 \times 10^3$.

11. Effectuer les opérations suivantes en arithmétique flottante à 3 chiffres.

    a) $\pi(1/\pi)$

    b) $2136\,(9993 + 0{,}004\,567)$

    c) $(1{,}235)^4$

    d) $10\,200 + 341$

    e) $(10\,200 + 341) - 9800$

    f) $(125 \times 25) + (10 \times 2{,}5)$

12. Combien de nombres différents peut-on représenter en arithmétique flottante à 3 chiffres (base 10) si l'exposant l est compris entre -9 et 9?

13. Est-ce que $(x \div y)$ est équivalent à $(x \times (1 \div y))$ en arithmétique flottante?

14. On doit effectuer l'opération $1 - \cos x$ pour des valeurs de x voisines de 0. Expliquer ce qui risque de se produire du point de vue de l'arithmétique flottante et proposer une solution de rechange.

15. Donner une façon d'évaluer les expressions suivantes qui permette d'éviter le plus possible les erreurs dues à l'arithmétique flottante.

    a) $\cos^2\theta - \sin^2\theta$ pour des valeurs de θ autour de $\pi/4$

    b) $p(2)$, où $p(x) = 1 - 2x + 3x^2 - 4x^3$

    c) $\displaystyle\sum_{i=1}^{100} \frac{1}{i^2}$

16. La série divergente:
$$\sum_{n=1}^{\infty} \frac{1}{n}$$
devient convergente en arithmétique flottante. Expliquer brièvement pourquoi.

17. Démontrer que l'erreur relative liée à la multiplication et à la division de deux nombres est la somme des erreurs relatives liées à chacun des nombres.

18. Effectuer les développements de Taylor suivants à l'ordre demandé. Utiliser la forme de l'équation 1.28. Donner l'expression analytique du terme d'erreur. Donner également une borne supérieure de l'erreur lorsque c'est possible.

    a) $\cos(x)$ autour de $x_0 = 0$ (ordre 8)

    b) $\sin(x)$ autour de $x_0 = 0$ (ordre 9)

    c) $\arctan(x)$ autour de $x_0 = 0$ (ordre 5)

    d) $\cos(x)$ autour de $x_0 = \pi/2$ (ordre 7)

    e) $\sin(x)$ autour de $x_0 = \pi/2$ (ordre 8)

19. Évaluer les erreurs commises dans l'évaluation des fonctions suivantes. Tous les chiffres fournis sont significatifs. Indiquer le nombre de chiffres significatifs du résultat.

    a) $\ln(x)$ en $x^* = 2{,}01$

    b) $\arctan(x)$ en $x^* = 1{,}0100$

    c) x^8 en $x^* = 1{,}123$

    d) $(\sin(x))^2$ en $x^* = 0{,}11$

20. Évaluer les erreurs commises dans l'évaluation des fonctions de plusieurs variables suivantes. Tous les chiffres fournis sont significatifs. Indiquer le nombre de chiffres significatifs du résultat.

    a) x^2y^3 en $x^* = 12{,}1,\ y^* = 3{,}721$

    b) $-xyz$ en $x^* = 1{,}260,\ y^* = 0.5 \times 10^{-3},\ z^* = 12{,}93$

21. À l'aide d'une méthode numérique, on a évalué la dérivée d'une fonction pour deux valeurs de h.

h	$f'(x_0)$	Erreur absolue
0,1	25,3121	0,0004
0,05	25,312 475	0,000 025

    a) Donner le nombre de chiffres significatifs de chaque approximation et arrondir au dernier chiffre significatif.

    b) Quel est l'ordre de la méthode de différentiation numérique utilisée.

22. a) Calculer le développement de Taylor d'ordre 5, c'est-à-dire dont le terme d'erreur est de type $O(h^5)$, de la fonction $f(x) = \ln(x)$ autour de $x_0 = 1$. Donner l'expression analytique du terme d'erreur.

    b) À l'aide de ce développement, donner une approximation de $\ln(1{,}1)$. Par comparaison avec la valeur exacte ($\ln 1{,}1 = 0{,}095\,310\,179\,8$), donner le nombre de chiffres significatifs de l'approximation.

    c) Par quel facteur approximatif l'erreur obtenue en b) serait-elle réduite si on évaluait $\ln(1{,}025)$ au moyen du développement de Taylor obtenu en a)? (Ne pas faire les calculs.)

23. En se servant d'un développement de Taylor de la fonction $\arctan x$ autour de $x_0 = 0$, on a obtenu les résultats suivants:

$$\arctan(0{,}4) = 0{,}380\,714\,667 \quad (\text{erreur absolue } = 0{,}208\,29 \times 10^{-3})$$
$$\arctan(0{,}1) = 0{,}099\,668\,667 \quad (\text{erreur absolue } = 0{,}1418 \times 10^{-7})$$

Quel était l'ordre du développement de Taylor utilisé?

24. a) Obtenir le développement de Taylor autour de $x_0 = 0$ de la fonction:

$$f(x) = \frac{1}{1-x}$$

b) Poser $x = -t^2$ dans le développement en a) et obtenir le développement de Taylor de:

$$g(t) = \frac{1}{1+t^2}$$

c) Intégrer l'expression obtenue en b) et obtenir le développement de Taylor d'$\arctan t$.

d) Utiliser l'expression obtenue en a) et obtenir le développement de Taylor de $\ln(1+x)$. (Remplacer x par $-x$ en premier lieu.)

25. La fonction d'erreur $f(x)$ est définie par:

$$f(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

Pour en obtenir le développement de Taylor, on peut suivre les étapes suivantes:

a) Obtenir le développement de Taylor de e^{-x}. (Limiter le développement aux 6 premiers termes.)

b) Déduire de a) le développement de Taylor de e^{-t^2}.

c) Déduire de b) le développement de Taylor de $f(x)$.

d) Donner une approximation de $f(1)$ en utilisant les 4 premiers termes de son développement de Taylor.

e) Quel est l'ordre de l'approximation obtenue en d)?

f) Donner le nombre de chiffres significatifs de l'approximation obtenue en d) en la comparant avec la valeur exacte $f(1) = 0{,}842\,701$.

Chapitre 2

Équations non linéaires

2.1 Introduction

Le numéricien est souvent confronté à la résolution d'équations algébriques de la forme:

$$f(x) = 0 \tag{2.1}$$

et ce dans toutes sortes de contextes. Introduisons dès maintenant la terminologie qui nous sera utile pour traiter ce problème.

Définition 2.1
Une valenr de x solution de $f(x) = 0$ est appelée une *racine* ou un *zéro* de la fonction $f(x)$ et est notée r.

Nous avons tous appris au secondaire comment résoudre l'équation du second degré:

$$ax^2 + bx + c = 0$$

dont les deux racines sont:

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Certains ont également vu comment calculer les racines d'une équation du troisième ordre et se souviennent que la formule est beaucoup plus complexe. On peut aussi obtenir une formule générale pour le quatrième degré. Par contre, la plupart des étudiants ignorent qu'il n'existe pas de formule permettant de trouver les racines des polynômes de degré plus grand ou égal

à 5. Non pas que les mathématiciens ne l'aient pas encore trouvée, mais Galois[1] a démontré que cette formule n'existe pas.

Puisqu'il n'existe pas de formule géuérale pour des fonctions aussi simples que des polynômes, il est peu probable que l'on puisse résoudre analytiquement l'équation 2.1 dans tous les cas qui nous intéressent. Il faudra donc recourir aux méthodes numériques. Dans ce qui suit, nous présentons plusieurs techniques de résolution, chacune ayant ses avantages et ses inconvénients. Nous tâcherons de mettre en évidence ces avantages et inconvénients de façon à tirer le meilleur parti de chacune des méthodes proposées.

Il faudra également se souvenir des enseignements du chapitre précédent pour éviter de développer des algorithmes numériquement instables.

2.2 Méthode de la bissection

La méthode de la bissection repose sur une idée toute simple, à savoir qu'en général, de part et d'autre d'une solntion de l'équation 2.1, une fonction continue $f(x)$ change de signe et passe du positif au négatif ou vice versa (voir la figure 2.1). De toute évidence, ce n'est pas toujours le cas puisque la fonction $f(x)$ peut aussi être tangente à l'axe des x (voir la figure 2.2).

Supposons pour l'instant qu'il y ait effectivement un changement de signe autour d'une racine r de $f(x)$. Nous nous occuperons des cas pathologiques un peu plus tard. Soit $[x_1, x_2]$, un intervalle ayant un changement de signe, c'est-à-dire:

$$f(x_1) \times f(x_2) < 0 \tag{2.2}$$

On pose alors:

$$x_m = \frac{x_1 + x_2}{2}$$

qui est bien sûr le point milieu de l'intervalle. Il suffit alors de déterminer, entre les intervalles $[x_1, x_m]$ et $[x_m, x_2]$, celui qui possède encore un changement de signe. La racine se trouvera forcément dans cet intervalle. À la première itération de la figure 2.1, ce serait l'intervalle $[x_m, x_2]$, tandis qu'à la deuxième itération ce serait $[x_1, x_m]$. Cela nous amène à l'algorithme suivant.

[1]Le mathématicien Évariste Galois (1811-1832) fut tué dans un duel à l'âge de 21 ans, non sans avoir eu le temps d'apporter une contribution considérable à la théorie des groupes.

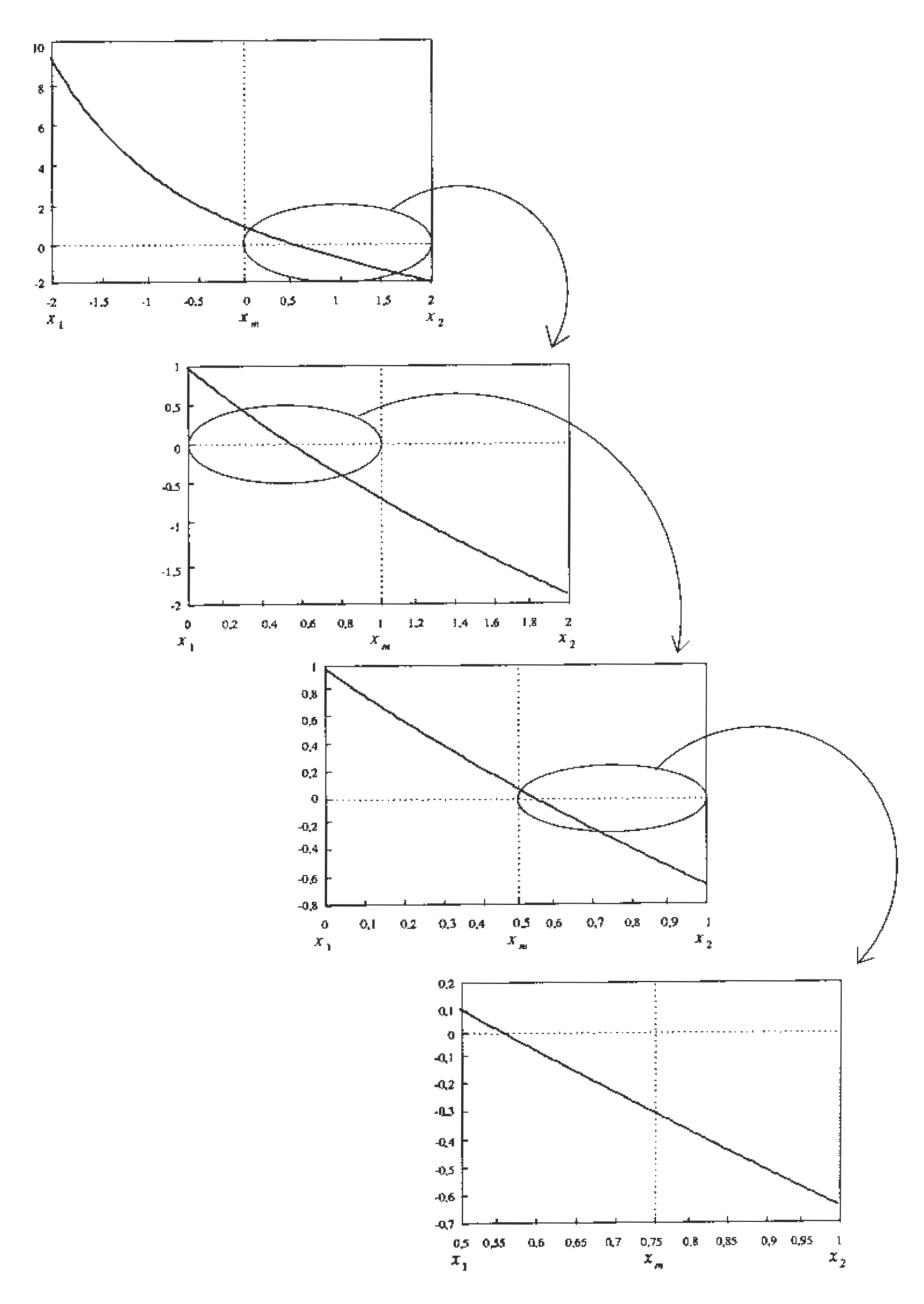


Figure 2.1: Méthode de la bissection: $f(x) = e^{-x} - x$

Algorithme 2.1: Algorithme de la bissection

1. Étant donné un intervalle $[x_1, x_2]$ pour lequel $f(x)$ possède un changement de signe

2. Étant donné ϵ, le critère d'arrêt, et N, le nombre maximal d'itérations

3. Poser:
$$x_m = \frac{x_1 + x_2}{2}$$

4. Si $\dfrac{|x_2 - x_1|}{2|x_m|} < \epsilon$:
   - convergence atteinte
   - écrire la racine x_m
   - écrire $f(x_m)$
   - arrêt

5. Écrire $x_1, x_2, x_m, f(x_1), f(x_2), f(x_m)$

6. Si $f(x_1) \times f(x_m) < 0$, alors $x_2 = x_m$

7. Si $f(x_m) \times f(x_2) < 0$, alors $x_1 = x_m$

8. Si le nombre maximal d'itérations N est atteint:
   - convergence non atteinte en N itérations
   - arrêt

9. Retour à l'étape 3 □

L'expression:
$$\frac{|x_2 - x_1|}{2|x_m|}$$
est une approximation de l'erreur relative. En effet, à l'étape 3 de l'algorithme de la bissection, la racine recherchée est soit dans l'intervalle $[x_1, x_m]$ ou dans l'intervalle $[x_m, x_2]$, qui sont tous deux de longueur:
$$\frac{x_2 - x_1}{2}$$

ce qui constitue une borne supérieure de l'erreur absolue. En divisant par x_m, on obtient une approximation assez fiable de l'erreur relative.

Remarque 2.1
Dans l'algorithme précédent, il faut prendre garde au cas où la racine recherchée est 0. Il y a alors risque de division par 0 au cours de l'évaluation de l'erreur relative. Ce cas est toutefois rare en pratique. □

Remarque 2.2
Il est parfois utile d'introduire un test d'arrêt sur la valeur de $f(x)$, qui doit tendre également vers 0. □

Exemple 2.1
La fonction $f(x) = x^3 + x^2 - 3x - 3$ possède un zéro dans l'intervalle $[1, 2]$. En effet:

$$f(1) \times f(2) = -4{,}0 \times 3{,}0 = -12{,}0 < 0$$

On a alors $x_m = 1{,}5$ et $f(1{,}5) = -1{,}875$. L'intervalle $[1{,}5\,,\,2]$ possède encore un changement de signe, ce qui n'est pas le cas pour l'intervalle $[1,\,1{,}5]$. Le nouvel intervalle de travail est donc $[1{,}5\,,\,2]$, dont le point milieu est $x_m = 1{,}75$. Puisque $f(1{,}75) = 0{,}171\,87$, on prendra l'intervalle $[1{,}5\,,\,1{,}75]$ et ainsi de suite. Le tableau suivant résume les résultats.

x_1	x_2	x_m	$f(x_1)$	$f(x_2)$	$f(x_m)$	Erreur absolue liée à x_m
1,0	2,0	1,5	−4,0	3,0	−1,875	0,5
1,5	2,0	1,75	−1,875	3,0	+0,171 87	0,25
1,5	1,75	1,625	−1,875	0,171 87	−0,943 35	0,125
1,625	1,75	1,6875	−0,943 35	0,171 87	−0,409 42	0,0625
1,6875	1,75	1,718 75	−0,409 42	0,171 87	−0,124 78	0,031 25

● ● ● ● ●

On remarque aisément que la longueur de l'intervalle entourant la racine est divisée par deux à chaque itération. Cette constatation permet de déterminer à l'avance le nombre d'itérations nécessaires pour obtenir une certaine

erreur absolue Δr sur la racine r. Soit $L = x_2 - x_1$, la longueur de l'intervalle de départ. Après une itération, le nouvel intervalle est de longueur $L/2$ et après n itérations la longueur de l'intervalle est:

$$\frac{L}{2^n}$$

Si on veut connaître la valeur de n nécessaire pour avoir:

$$\frac{L}{2^n} < \Delta r$$

il suffit de résoudre cette équation en fonction de n et on trouve la condition:

$$n > \frac{\ln(\frac{L}{\Delta r})}{\ln 2} \tag{2.3}$$

Il est clair que, sur le plan pratique, on doit prendre pour valeur de n le plus petit entier vérifiant cette condition.

Exemple 2.2
Dans l'exemple précédent, $L = 2{,}0 - 1{,}0$. Si on veut une erreur absolue plus petite que $0{,}5 \times 10^{-2}$, ce qui revient à s'assurer que le chiffre des centièmes est significatif, il faut effectuer au moins:

$$\frac{\ln(\frac{1{,}0}{0{,}5\times 10^{-2}})}{\ln 2} = 7{,}64 \text{ itérations}$$

On fera donc 8 itérations pour s'assurer de cette précision. On peut aisément vérifier qu'après 8 itérations l'erreur maximale liée à x_m est de 0,003 906 25 et que la véritable erreur est 0,001 582.

● ● ● ● ●

Exemple 2.3
On souhaite calculer $\sqrt{2}$ avec une calculatrice dotée seulement des 4 opérations élémentaires. Cela revient à résoudre:

$$x^2 - 2 = 0$$

Cette fonction présente un changement de signe dans l'intervalle $[1, 2]$. L'algorithme de la bissection donne les résultats suivants avec $\epsilon = 10^{-3}$.

x_1	x_2	x_m	$f(x_1)$	$f(x_2)$	$f(x_m)$	$(x_m)_2$
1,0	2,0	1,5	−1,0	2,0	+0,25	1,1
1,0	1,5	1,25	−1,0	0,25	−0,4375	1,01
1,25	1,5	1,375	−0,4375	0,25	−0,1094	1,011
1,375	1,5	1,4375	−0,1094	0,25	+0,066 41	1,0111
1,375	1,4375	1,4062	−0,1094	0,066 41	−0,022 46	1,011 01
1,4062	1,4375	1,4219	−0,022 46	0,066 41	+0,021 73	1,011 011
1,4062	1,4219	1,4141	−0,022 46	0,021 73	−0,000 43	1,011 010 1
1,4141	1,4219	1,4180	−0,000 43	0,021 73	+0,010 64	⋮
1,4141	1,4180	1,4160	−0,000 43	0,010 64	+0,005 10	
1,4141	1,4160	1,4150	−0,000 43	0,005 10	+0,002 33	
1,4141	1,4150	1,4146	−0,000 43	0,002 33	+0,000 95	
1,4141	1,4146	1,4143	−0,000 43	0,000 95	+0,000 26	

On a arrondi à 5 chiffres les résultats de ce tableau. La racine trouvée est alors 1,414 184 57, ce qui se rapproche de la valeur exacte (à 10 chiffres significatifs) 1,414 213 562. Cet exemple est particulièrement intéressant du point de vue de la représentation binaire. En effet, l'intervalle de départ étant $[1, 2]$ et puisque les nombres 1 et 2 possèdent des représentations binaires exactes, chaque itération de la méthode de la bissection permet de fixer 1 bit de la représentation binaire de la racine. À la $(n+1)^e$ itération, on est assuré que les n premiers bits de la mantisse de x_m sont exacts. On peut constater ce phénomène à la dernière colonne du tableau, qui contient la représentation binaire de x_m.

● ● ● ● ●

Remarque 2.3

La convergence de la méthode de la bissection n'est pas très rapide, mais elle est sûre à partir du moment où on a un intervalle avec changement de signe. On parle alors de *méthode fermée*, car on travaille dans un intervalle fermé. C'est également le cas de la *méthode de la fausse position* (voir les exercices de fin de chapitre). Les méthodes des sections qui suivent sont dites *ouvertes* en ce sens qu'il n'y a pas d'intervalle à déterminer ayant un changement de signe. Au contraire des méthodes fermées, les méthodes ouvertes ne garantissent nullement la convergence, mais elles possèdent d'autres avantages. □

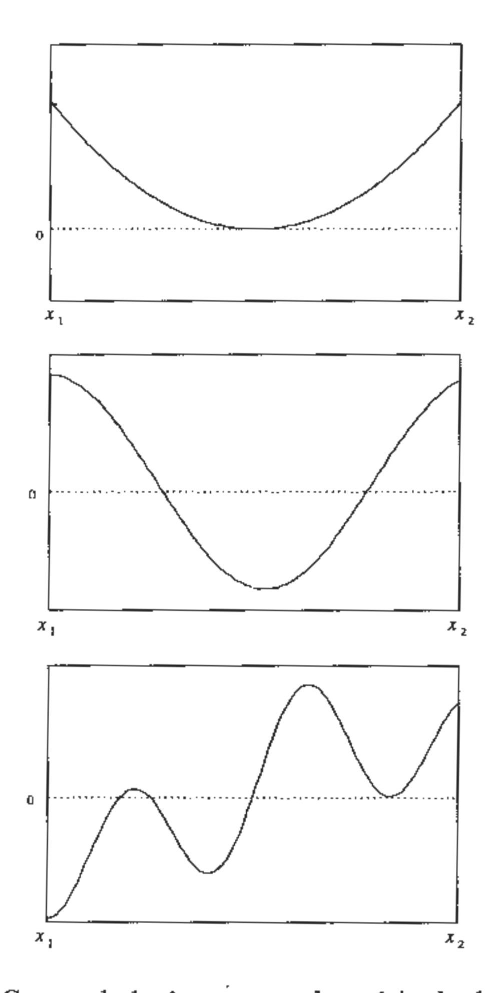


Figure 2.2: Cas pathologiques pour la méthode de la bissection

Remarque 2.4
Il existe des cas où la méthode de la bissection achoppe. La figure 2.2 illustre certains de ces cas. La première situation critique est celle où la fonction $f(x)$ est tangente à l'axe des x et ne présente donc pas de changement de signe. La bissection ne peut alors s'appliquer. Il y a anssi celle où deux racines (ou un nombre pair de racines) sont présentes dans l'intervalle de départ; en ce cas, il n'y a toujours pas de changement de signe. Enfin, si l'intervalle de départ contient un nombre impair de racines, $f(x)$ change de signe mais l'algorithme peut avoir des difficultés à choisir parmi ces racines. On peut assez facilement éviter ces écueils en illustrant graphiquement la fonction $f(x)$ dans l'intervalle d'intérêt. □

2.3 Méthodes des points fixes

Avant d'aborder les méthodes des points fixes, il importe de définir ce qu'est un point fixe d'une fonction.

Définition 2.2
Un *point fixe* d'une fonction $g(x)$ est une valeur de x qui reste invariante pour cette fonction, c'est-à-dire toute solution de:

$$x = g(x) \tag{2.4}$$

est un point fixe de la fonction $g(x)$.

Il existe nn algorithme très simple permettant de déterminer des points fixes. Il snffit en effet d'effectuer les itérations de la façon suivante:

$$\left\{ \begin{array}{lcl} x_0 & & \text{donné} \\ x_{n+1} & = & g(x_n) \end{array} \right. \tag{2.5}$$

à partir d'une valeur estimée initiale x_0. L'intérêt de cet algorithme réside dans sa généralité et dans la relative facilité avec laquelle on peut en faire l'analyse de convergence. Il en résulte l'algorithme plus complet suivant.

Algorithme 2.2: Algorithme des points fixes

1. Étant donné ϵ, un critère d'arrêt
2. Étant donné N, le nombre maximal d'itérations
3. Étant donné x_0, une valeur estimée initiale du point fixe
4. Effectuer $x_{n+1} = g(x_n)$
5. Si $\dfrac{|x_{n+1} - x_n|}{|x_{n+1}|} < \epsilon$:
   - convergence atteinte
   - écrire la solution x_{n+1}
   - arrêt
6. Si le nombre maximal d'itérations N est atteint:
   - convergence non atteinte en N itérations
   - arrêt
7. Retour à l'étape 4 □

On peut résoudre des équations non linéaires de la forme $f(x) = 0$ en utilisant l'algorithme des points fixes. Il suffit pour ce faire de transformer l'équation $f(x) = 0$ en un *problème équivalent* de la forme $x = g(x)$. L'ennui, c'est qu'il y a une infinité de façons différentes de le faire. Nous verrons que certains choix donnent lieu à des algorithmes convergents et d'autres pas.

Exemple 2.4

Commençons par un exemple simple. On cherche à résoudre l'équation du second degré $x^2 - 2x - 3 = 0$. Il n'est pas nécessaire de recourir aux méthodes numériques pour résoudre ce problème, dont les deux solutions sont $r_1 = 3$ et $r_2 = -1$. Cet exemple permet cependant de mieux comprendre ce qui se passe lorsqu'on utilise l'algorithme des points fixes. Pnisqu'il y a une infinité de façons différentes de transformer cette équation sous la forme $x = g(x)$, nous en choisissons trois au hasard. Vous pouvez bien sûr recourir à d'autres.

$$
\begin{aligned}
x &= \sqrt{2x+3} &&= g_1(x) \quad \text{(en isolant } x^2\text{)} \\
x &= \frac{3}{x-2} &&= g_2(x) \quad \text{(en écrivant } x(x-2)-3=0\text{)} \\
x &= \frac{x^2-3}{2} &&= g_3(x) \quad \text{(en isolant le } x \text{ de } -2x\text{)}
\end{aligned}
\tag{2.6}
$$

Si on applique l'algorithme des points fixes à chacune des fonctions $g_i(x)$ en partant de $x_0 = 4$, on obtient pour $g_1(x)$:

$$
\begin{aligned}
x_1 &= g_1(4) &&= \sqrt{2 \times 4 + 3} &&= 3{,}316\,6248 \\
x_2 &= g_1(3{,}316\,6248) &&= \sqrt{2 \times 3{,}316\,6248 + 3} &&= 3{,}103\,7477 \\
x_3 &= g_1(3{,}103\,7477) &&= \sqrt{2 \times 3{,}103\,7477 + 3} &&= 3{,}034\,3855 \\
x_4 &= g_1(3{,}034\,3855) &&= \sqrt{2 \times 3{,}034\,3855 + 3} &&= 3{,}011\,4402 \\
&\vdots && \vdots && \vdots \\
x_{10} &= g_1(3{,}000\,0470) &&= \sqrt{2 \times 3{,}000\,0470 + 3} &&= 3{,}000\,0157
\end{aligned}
$$

L'algorithme semble donc converger vers la racine $r_1 = 3$. Reprenons l'exercice avec $g_2(x)$, toujours en partant de $x_0 = 4$:

$$
\begin{aligned}
x_1 &= g_2(4) &&= \frac{3}{4-2} &&= 1{,}5 \\
x_2 &= g_2(1{,}5) &&= \frac{3}{1{,}5-2} &&= -6{,}0 \\
x_3 &= g_2(-6{,}0) &&= \frac{3}{-6{,}0-2} &&= -0{,}375 \\
x_4 &= g_2(-0{,}375) &&= \frac{3}{-0{,}375-2} &&= -1{,}263\,1579 \\
&\vdots && \vdots && \vdots \\
x_{10} &= g_2(-0{,}998\,9841) &&= \frac{3}{-0{,}998\,9841-2} &&= -1{,}000\,3387
\end{aligned}
$$

On remarque que, contrairement au cas précédent, les itérations convergent vers la racine $r_2 = -1$ en ignorant la racine $r_1 = 3$. En dernier lieu, essayons

l'algorithme avec la fonction $g_3(x)$:

$$\begin{aligned}
x_1 &= g_3(4) &&= \frac{(4)^2-3}{2} &&= 6{,}5 \\
x_2 &= g_3(6{,}5) &&= \frac{(6{,}5)^2-3}{2} &&= 19{,}625 \\
x_3 &= g_3(19{,}625) &&= \frac{(19{,}625)^2-3}{2} &&= 191{,}0703 \\
x_4 &= g_3(191{,}0703) &&= \frac{(191{,}0703)^2-3}{2} &&= 18\,252{,}43 \\
& \vdots && \vdots && \vdots
\end{aligned}$$

Visiblement, les itérations tendent vers l'infini et aucune des deux solutions possibles ne sera atteinte.

Cet exemple montre clairement qne l'algorithme des points fixes, selon le choix de la fonction itérative $g(x)$, converge vers l'une ou l'autre des racines et peut même diverger complètement dans certains cas. Il faut donc une analyse plus fine afin de déterminer dans quelles conditions la méthode des points fixes est convergente.

●●●●●

2.3.1 Convergence de la méthode des points fixes

Nons nous intéressons dans cette section au comportement de la méthode des points fixes pour la résolution de l'équation $f(x) = 0$. On a d'abord transformé cette équation sous la forme équivalente $x = g(x)$. Soit r, une valeur qui est à la fois une racine de $f(x)$ et un point fixe de la fonction $g(x)$, c'est-à-dire qui vérifie $f(r) = 0$ et:

$$r = g(r) \tag{2.7}$$

On définit l'erreur à l'étape n comme étant:

$$e_n = x_n - r$$

On cherche à déterminer sous quelles conditions l'algorithme des points fixes converge vers la racine r. Ce sera bien sûr le cas si l'erreur e_n tend vers

0 lorsque n devient grand. Il est intéressant de suivre le comportement de l'erreur au fil des itérations. On a en vertu des relations 2.5 et 2.7:

$$e_{n+1} = x_{n+1} - r = g(x_n) - g(r) \tag{2.8}$$

On constate aisément que:

$$x_n = r + (x_n - r) = r + e_n$$

et on peut alors utiliser un développement de Taylor de la fonction $g(x)$ autour de la racine r. La relation 2.8 devient alors:

$$\begin{aligned} e_{n+1} &= g(r + e_n) - g(r) \\ &= \left(g(r) + g'(r)e_n + \frac{g''(r)e_n^2}{2!} + \frac{g'''(r)e_n^3}{3!} + \cdots \right) - g(r) \end{aligned}$$

On en conclut que:

$$e_{n+1} = g'(r)e_n + \frac{g''(r)e_n^2}{2} + \frac{g'''(r)e_n^3}{3!} + \cdots \tag{2.9}$$

L'étnde de la relation 2.9 est fondamentale pour la compréhension des méthodes de points fixes. Au voisinage de la racine r, le premier terme non nul de l'expression de droite sera déterminant pour la convergence.

Selon l'équation 2.9, si $g'(r) \neq 0$ et si on néglige les termes d'ordre supérieur ou égal à 2 en e_n, on a:

$$e_{n+1} \simeq g'(r)e_n \tag{2.10}$$

On voit qne l'erreur à l'étape $(n+1)$ est directement proportionnelle à l'erreur à l'étape n. L'erreur ne pourra donc diminuer qne si:

$$|g'(r)| < 1 \tag{2.11}$$

La condition 2.11 est une condition nécessaire de convergence d'une méthode de points fixes. On remarque également que le signe de $g'(r)$ a une influence sur la convergence. En effet, si:

$$-1 < g'(r) < 0$$

l'errenr changera de signe à chaque itération en vertu de l'équation 2.10 et les valeurs de x_n oscilleront de part et d'autre de r. La convergence n'en sera pas moins assurée.

La relation 2.10 donne de plus la vitesse à laquelle l'erreur diminue. En effet, plns $g'(r)$ est petit, plus l'erreur diminue vite et donc plus la convergence est rapide. Cela nous amène à la définition suivante.

Définition 2.3
Le *taux de convergence* d'une méthode de points fixes est donné par $|g'(r)|$.

Plus le taux de convergence est petit, plus la convergence est rapide. Le cas limite est celui où $g'(r) = 0$. Dans ce cas, on déduit de l'équation 2.9 que l'erreur e_{n+1} est proportionnelle à e_n^2. Cela nous amène à une autre définition.

Définition 2.4
On dit qn'nne méthode de points fixes *converge à l'ordre p* si:

$$|e_{n+1}| \simeq C\,|e_n|^p \tag{2.12}$$

où C est nne constante. La convergence d'ordre 1 est également dite *linéaire*, tandis que celle d'ordre 2 est dite *quadratique*.

Remarque 2.5
Si $|g'(r)| < 1$ et $|g'(r)| \neq 0$, la méthode de points fixes converge à l'ordre 1. Si $|g'(r)| = 0$ et $|g''(r)| \neq 0$, on a une convergence quadratique; si $|g'(r)| = |g''(r)| = 0$ et $|g'''(r)| \neq 0$, la convergence est d'ordre 3; et ainsi de suite. □

Remarque 2.6
La convergence d'une méthode de points fixes est également assujettie au choix de la valenr initiale x_0. En effet, un mauvais choix de x_0 peut résulter en un algorithme divergent même si la condition 2.11 est respectée. □

Cela nons amène à définir le bassin d'attraction d'nne racine r.

Définition 2.5
Le *bassin d'attraction* de la racine r pour la méthode de points fixes $x_{n+1} = g(x_n)$ est l'ensemble des valeurs initiales x_0 pour lesquelles x_n tend vers r lorsque n tend vers l'infini.

En d'autres termes, le bassin d'attraction de r comprend tous les points x_0 pour lesquels la méthode de points fixes converge vers r. Pour s'assurer de la convergence, il faut donc choisir x_0 dans le bassin d'attraction de r. Intuitivement, on choisit x_0 aussi près que possible de r en utilisant par exemple une méthode graphique. Il faut aussi se souvenir que les problèmes rencontrés proviennent le plus souvent de l'ingénierie et que le numéricien doit ntiliser ses connaissances pour estimer x_0. Par exemple, si la racine que l'on cherche correspond à une longueur ou à une concentration, il serait peu raisonnable de prendre une valeur négative de x_0. Très souvent, le simple bon sens permet de choisir x_0 avec succès.

Définition 2.6
Un point fixe r de la fonction $g(x)$ est dit *attractif* si:

$$|g'(r)| < 1$$

et *répulsif* si:

$$|g'(r)| > 1$$

Le cas où $|g'(r)| = 1$ est indéterminé.

Exemple 2.5
Considérons la fonction $g(x) = x^2$ qui possède les points fixes $x = 0$ et $x = 1$. Ce dernier est répulsif, car la dérivée de $g(x)$ $(2x)$ vaut 2 en $x = 1$. Le seul point fixe intéressant est donc $x = 0$. La méthode des points fixes engendre, à partir de la valeur initiale x_0, la suite:

$$x_0, \ x_0^2, \ x_0^4, \ x_0^8, \ x_0^{16}, \ x_0^{32} \ \cdots$$

Cette suite convergera vers 0 seulement si $x_0 \in \,]-1, 1[$. Ce dernier intervalle

constitue donc le bassin d'attraction de ce point fixe. Toute valeur de x_0 choisie à l'extérieur de cet intervalle résultera en un algorithme divergent.

● ● ● ● ●

Remarque 2.7
Dans le cas d'un point fixe répulsif, le bassin d'attraction se réduit à peu de choses, le plus souvent à l'ensemble $\{r\}$ constitué d'un seul point. □

Le résultat suivant permet dans certains cas de s'assurer de la convergence (voir Burden et Faires, réf. [2]).

Théorème 2.1
Soit $g(x)$, une fonction continue dans l'intervalle $I = [a, b]$ et telle que $g(x) \in I$ pour tout x dans I. Si de plus $g'(x)$ existe et si:

$$|g'(x)| \leq k < 1$$

pour tout x dans l'intervalle ouvert (a, b), alors tous les points x_0 de l'intervalle I appartiennent au bassin d'attraction de l'*unique* point fixe r de I. □

Remarque 2.8
Il est possible que la méthode des points fixes converge dans le cas où:

$$|g'(r)| = 1$$

Il s'agit d'un cas limite intéressant. Nous verrons plus loin un exemple de cette situation. La convergence dans ce cas est au mieux extrêmement lente, car le taux de convergence est près de 1. □

Exemple 2.6
Revenons aux trois fonctions $g_i(x)$ de l'exemple précédent. On veut s'assurer que la condition 2.11 est vérifiée à l'une ou l'autre des racines $r_1 = 3$ et

$r_2 = -1$. On doit d'abord calculer les dérivées:

$$
\begin{aligned}
g_1'(x) &= \frac{1}{\sqrt{2x+3}} \\
g_2'(x) &= \frac{-3}{(x-2)^2} \\
g_3'(x) &= x
\end{aligned}
$$

Les résultats sont compilés dans le tableau suivant.

	$r_1 = 3$	$r_2 = -1$
$g_1'(r)$	0,333 33	1
$g_2'(r)$	-3	$-0{,}333\,33$
$g_3'(r)$	3	-1

Ce tableau aide à comprendre les résultats obtenus précédemment. La méthode de points fixes appliquée à $g_1(x)$ a convergé vers $r_1 = 3$, puisque $g_1'(3) < 1$. De même, avec $g_2(x)$, la méthode de points fixes ne peut converger vers $r_1 = 3$, car la dérivée de $g_2(x)$ en ce point est plus grande que 1. Les itérations ignorent r_1 et convergent vers r_2, où la valeur de la dérivée est inférieure à 1.

Enfin, la fonction $g_3(x)$ a également une dérivée plus grande que 1 en r_1. L'analyse de la convergence autour de $r_2 = -1$ est plus subtile. En effet, puisque $g_3'(x) = x$, on constate que la valeur absolue de la dérivée est inférieure à 1 à droite de r_2 et supérieure à 1 à gauche de r_2. De plus, cette dérivée est négative, ce qui signifie que la méthode des points fixes oscillera de part et d'autre de la racine. À une itération, la pente $g'(x_n)$ sera inférieure à 1 et à l'itération suivante la pente $g'(x_{n+1})$ sera supérieure à 1 en valeur absolue. On en conclut que l'algorithme de points fixes s'approchera légèrement de r_2 à une itération et s'en éloignera à la suivante. En un mot, l'algorithme piétinera. On peut vérifier ce raisonnement en effectuant les itérations à partir de $x_0 = -0{,}95$. On obtient après 10 000 itérations la valeur $x_{10000} = -0{,}986\,36$, ce qui signifie que la convergence est extrêmement lente.

● ● ● ● ●

Exemple 2.7
Considérons la fonction $g(x) = x^2 + 1/4$ dont l'unique point fixe est 1/2. On a bien sûr $g'(x) = 2x$, qui vaut précisément 1 en 1/2. En partant de $x_0 = 1/4$, on obtient la valeur 0,499 009 5 après 1000 itérations et donc une convergence très lente. Cela s'explique par le fait que la dérivée de $g(x)$ est légèrement inférieure à 1 pour les valeurs de x inférieures à 1/2 et que les résultats des itérations restent toujours inférieurs à 1/2. Par contre, en partant de $x_0 = 0{,}51$, l'algorithme diverge violemment après une centaine d'itérations. On constate aisément que la dérivée de $g(x)$ est supérieure à 1 pour les valeurs de x supérieures à 1/2.

• • • • •

2.3.2 Interprétation géométrique

L'algorithme de points fixes possède une interprétation géométrique très élégante qui permet d'illustrer la convergence ou la divergence. La figure 2.3 présente les différents cas possibles: $0 < g'(r) < 1$, $-1 < g'(r) < 0$ et $g'(r) > 1$. On peut interpréter cette figure de la manière suivante. Les courbes $y = x$ et $y = g(x)$ sont représentées et les points fixes sont bien entendu à l'intersection de ces deux courbes. À partir de la valeur initiale x_0, on se rend sur la courbe $y = g(x)$ au point $(x_0, g(x_0))$ et de là sur la droite $y = x$ au point $(g(x_0), g(x_0))$, qui est en fait (x_1, x_1). On recommence le même processus à partir de x_1 pour se rendre à $(x_1, g(x_1))$ et de là sur la droite $y = x$ au point $(g(x_1), g(x_1)) = (x_2, x_2)$. On répète ce trajet jusqu'à la convergence (ou la divergence) de l'algorithme.

On voit immédiatement la différence de comportement entre les cas convergents $0 < g'(r) < 1$ et $-1 < g'(r) < 0$. Bien que les itérations oscillent de part et d'autre de la racine lorsque la pente est négative, la convergence n'en est pas moins assurée. Par contre, lorsque la pente est supérieure à 1, les itérations s'éloignent de la racine recherchée. On obtiendrait un résultat similaire dans le cas où $g'(r) < -1$; les itérations s'éloigneraient de la racine en oscillant de part et d'autre de la racine.

Nous terminons cette section par un dernier exemple qui illustre la convergence généralement linéaire des méthodes de points fixes.

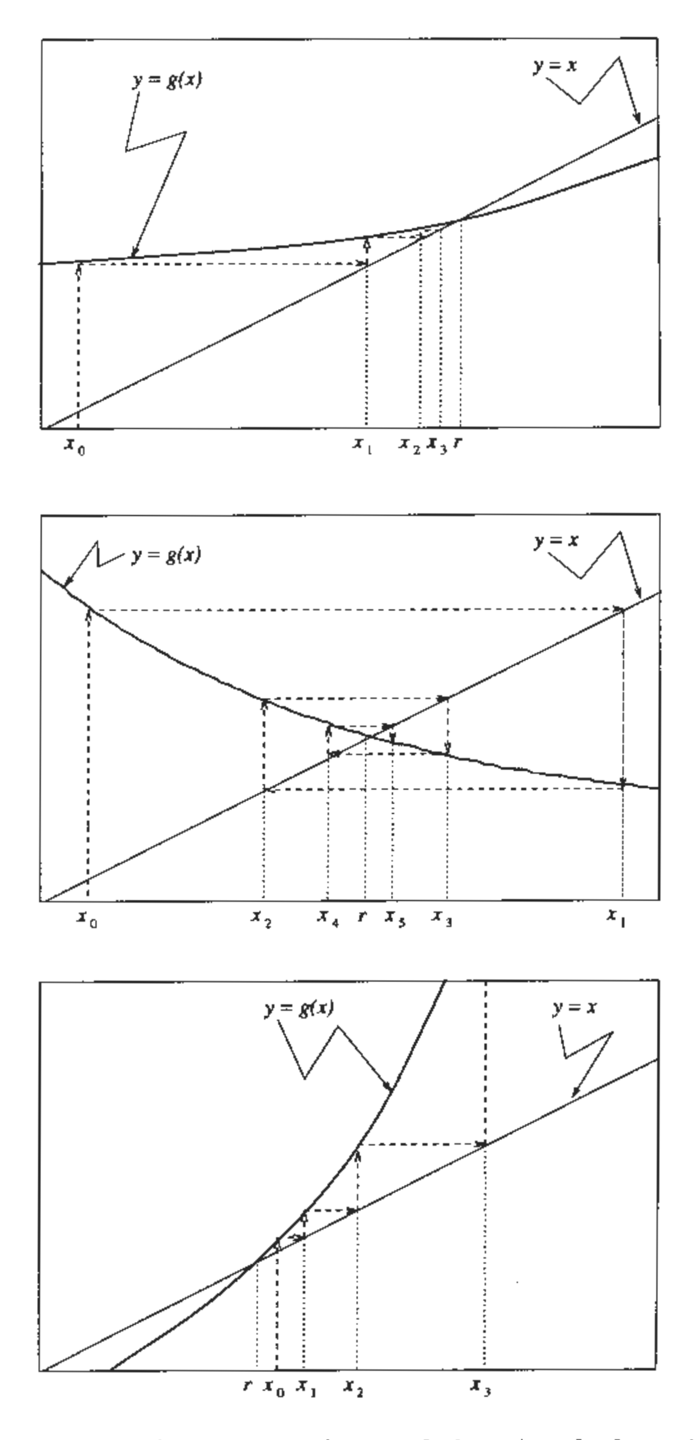


Figure 2.3: Interprétation géométrique de la méthode des points fixes

Exemple 2.8
On considère la résolution de $e^{-x} - x = 0$, que l'on transforme en un problème de points fixes $x = e^{-x}$. En partant de $x_0 = 0$ et en posant $e_n = x_n - r$, l'erreur à l'étape n, on obtient le tableau suivant.

n	x_n	$\lvert e_n \rvert$	$\left\lvert \frac{e_n}{e_{n-1}} \right\rvert$
1	1,000 0000	$0{,}4328 \times 10^{+0}$	—
2	0,367 8794	$0{,}1992 \times 10^{+0}$	0,4603
3	0,692 2006	$0{,}1250 \times 10^{+0}$	0,6276
4	0,500 4735	$0{,}6667 \times 10^{-1}$	0,5331
5	0,606 2435	$0{,}3910 \times 10^{-1}$	0,5864
6	0,545 3957	$0{,}2174 \times 10^{-1}$	0,5562
7	0,579 6123	$0{,}1246 \times 10^{-1}$	0,5733
⋮	⋮	⋮	⋮
14	0,566 9089	$0{,}2344 \times 10^{-3}$	0,5670
15	0,567 2762	$0{,}1329 \times 10^{-3}$	0,5672
⋮	⋮	⋮	⋮
35	0,567 1433	$\simeq 0$	0,5671

Pour remplir ce tableau, on a d'abord calculé la racine $r = 0{,}567\,143\,29$, ce qui a permis d'évaluer les erreurs par la suite. L'analyse de ce tableau illustre plusieurs points déjà discutés. En premier lieu, on constate la convergence vers la racine $r = 0{,}567\,143\,29$ puisque l'erreur e_n tend vers 0. Fait plus important encore, la troisième colonne converge vers environ 0,5671. Ce nombre n'est pas arbitraire. En effet, en vertu de la relation 2.10, ce ratio doit converger vers $|g'(r)|$, qui vaut dans cet exemple 0,567 14. On constate bien la convergence de ce ratio vers $|g'(r)|$ pour cet exemple.

● ● ● ● ●

2.3.3 Extrapolation d'Aitken

À partir d'une méthode de points fixes convergeant à l'ordre 1, on peut obtenir une méthode convergeant à l'ordre 2. Il suffit de remarquer que pour une méthode d'ordre 1:

$$e_2 \simeq g'(r)e_1$$

et

$$e_1 \simeq g'(r)e_0$$

On a alors immédiatement:

$$\frac{e_2}{e_1} \simeq \frac{e_1}{e_0}$$

c'est-à-dire

$$\frac{x_2 - r}{x_1 - r} \simeq \frac{x_1 - r}{x_0 - r}$$

En isolant r, on trouve facilement que:

$$r \simeq \frac{x_2 x_0 - x_1^2}{x_2 - 2x_1 + x_0}$$

qui est une formule numériquement instable. On lui préférera l'expression équivalente:

$$r \simeq = x_0 - \frac{(x_1 - x_0)^2}{x_2 - 2x_1 + x_0} \tag{2.13}$$

La relation 2.13 est dite *formule d'extrapolation d'Aitken* et permet d'obtenir à partir de x_0, x_1 et x_2 une meilleure approximation du point fixe r. Cela peut résulter en un algorithme qui accélère grandement la convergence d'une méthode de points fixes. C'est l'*algorithme de Steffenson.*

Algorithme 2.3: Algorithme de Steffenson

1. Étant donné ϵ, un critère d'arrêt
2. Étant donné N, le nombre maximal d'itérations
3. Étant donné x_0, une valeur estimée initiale du point fixe
4. Effectuer:
   - $x_1 = g(x_0)$
   - $x_2 = g(x_1)$
   - $x_e = x_0 - \dfrac{(x_1 - x_0)^2}{x_2 - 2x_1 + x_0}$
5. Si $\dfrac{|x_e - x_0|}{|x_e|} < \epsilon$:
   - convergence atteinte

- écrire la solution x_e
- arrêt

6. Si le nombre maximal d'itérations N est atteint:

- convergence non atteinte en N itérations
- arrêt

7. $x_0 = x_e$ et retour à l'étape 4 □

Exemple 2.9
Reprenons l'exemple précédent de la méthode des points fixes:

$$x_{n+1} = g(x_n) = e^{-x_n}$$

en partant de $x_0 = 0$. L'algorithme de Steffenson consiste à faire deux itérations de points fixes, à extrapoler pour obtenir x_e, à faire deux nouvelles itérations de points fixes à partir de x_e, à extrapoler à nouveau et ainsi de suite. On obtient dans ce cas:

$$\begin{aligned} x_1 &= e^0 &= 1{,}0 \\ x_2 &= e^{-1} &= 0{,}367\,8794 \end{aligned}$$

La valeur extrapolée est alors:

$$x_e = 0 - \frac{(1-0)^2}{0{,}367\,8794 - 2(1) + 0} = 0{,}612\,6998$$

À partir de cette nouvelle valeur, on fait deux itérations de points fixes:

$$x_1 = e^{-0{,}612\,6998} = 0{,}541\,8859$$

$$x_2 = e^{-0{,}541\,8859} = 0{,}581\,6503$$

La valeur extrapolée est alors:

$$\begin{aligned} x_e &= 0{,}612\,6998 - \frac{(0{,}541\,8859 - 0{,}612\,6998)^2}{0{,}581\,6503 - 2(0{,}541\,8859) + 0{,}612\,6998} \\ &= 0{,}567\,3509 \end{aligned}$$

En continuant ainsi, on obtient:

x_0	x_1	x_2	x_e
0,567 3509	0,567 0256	0,567 2101	0,567 1433
0,567 1433	0,567 1433	0,567 1433	0,567 1433

On remarque que la convergence est plus rapide avec l'algorithme de Steffenson qu'avec la méthode de points fixes dont elle est issue. Quatre itérations suffisent pour obtenir la même précision. On peut montrer en fait que la convergence est quadratique. On note toutefois que chaque itération de l'algorithme de Steffenson demande plus de calculs qu'une méthode de points fixes. Il y a un prix à payer pour obtenir une convergence quadratique.

●●●●●

2.4 Méthode de Newton

La méthode de Newton est l'une des méthodes les plus utilisées pour la résolution des équations non linéaires. Cette méthode possède également une belle interprétation géométrique. Nous commençons cependant par donner une première façon d'en obtenir l'algorithme, basée sur l'utilisation du développement de Taylor. Cette approche est également valable pour les systèmes d'équations non linéaires que nous verrons au chapitre 3.

Soit une équation à résoudre de la forme:

$$f(x) = 0$$

À partir d'une valeur initiale x_0 de la solution, on cherche une correction δx telle que:

$$0 = f(x_0 + \delta x)$$

En faisant un développement de Taylor autour de $x = x_0$, on trouve:

$$0 = f(x_0) + f'(x_0)\delta x + \frac{f''(x_0)(\delta x)^2}{2!} + \frac{f'''(x_0)(\delta x)^3}{3!} + \cdots$$

Il suffit maintenant de négliger les termes d'ordre supérieur ou égal à 2 en δx pour obtenir:

$$0 \simeq f(x_0) + f'(x_0)\delta x$$

On peut alors isoler la correction recherchée:

$$\delta x = -\frac{f(x_0)}{f'(x_0)}$$

La correction δx est en principe la quantité que l'on doit ajouter à x_0 pour annuler la fonction $f(x)$. Puisque nous avons négligé les termes d'ordre supérieur ou égal à 2 dans le développement de Taylor, cette correction n'est pas parfaite et on pose:

$$x_1 = x_0 + \delta x$$

On recommence le processus en cherchant à corriger x_1 d'une nouvelle quantité δx. On obtient alors l'algorithme suivant.

Algorithme 2.4: Algorithme de la méthode de Newton

1. Étant donné ϵ, un critère d'arrêt
2. Étant donné N, le nombre maximal d'itérations
3. Étant donné x_0, une valeur initiale de la solution
4. Effectuer: $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$
5. Si $\dfrac{|x_{n+1} - x_n|}{|x_{n+1}|} < \epsilon$:
   - convergence atteinte
   - écrire la solution x_{n+1}
   - arrêt
6. Si le nombre maximal d'itérations N est atteint:
   - convergence non atteinte en N itérations
   - arrêt
7. retour à l'étape 4 □

Remarque 2.9
L'algorithme de la méthode de Newton est un cas particulier de celui de la méthode des points fixes où:

$$g(x) = x - \frac{f(x)}{f'(x)} \qquad \square$$

Exemple 2.10
On cherche à résoudre l'équation $f(x) = e^{-x} - x = 0$. Pour utiliser la méthode de Newton, il faut obtenir la dérivée de cette fonction, qui est $f'(x) = -e^{-x} - 1$. L'algorithme se résume à:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{e^{-x_n} - x_n}{-e^{-x_n} - 1}$$

Les résultats sont compilés dans le tableau suivant à partir de $x_0 = 0$.

n	x_n	$\lvert e_n \rvert$	$\lvert \frac{e_n}{e_{n-1}} \rvert$
0	0,000 0000	$0{,}5671 \times 10^{+0}$	—
1	0,500 0000	$0{,}6714 \times 10^{-1}$	$0{,}1183 \times 10^{+0}$
2	0,566 3110	$0{,}8323 \times 10^{-3}$	$0{,}1239 \times 10^{-1}$
3	0,567 1432	$0{,}1250 \times 10^{-6}$	$0{,}1501 \times 10^{-3}$
4	0,567 1433	$0{,}4097 \times 10^{-9}$	$\simeq 0$

On remarque la convergence très rapide de cette méthode. Il suffit en effet pour s'en convaincre de comparer ces valeurs avec les résultats obtenus avec la méthode des points fixes pour le même problème. On note également que le nombre de chiffres significatifs double à chaque itération. Ce phénomène est caractéristique de la méthode de Newton et nous en verrons la raison au moment de l'analyse de convergence. La dernière colonne, qui converge vers 0, donne une indication à ce sujet. Cette colonne est censée converger vers $|g'(r)|$, qui est donc nul dans ce cas, ce qui semble indiquer que la convergence est quadratique.

●●●●●

2.4.1 Interprétation géométrique

La figure 2.4 permet de donner une interprétation géométrique assez simple de la méthode de Newton. Sur cette figure, on a représenté la fonction $f(x)$, la valeur initiale x_0 et le point $(x_0, f(x_0))$ qui est sur la courbe. La droite tangente à la courbe en ce point est de pente $f'(x_0)$ et a pour équation:

$$y = f(x_0) + f'(x_0)(x - x_0)$$

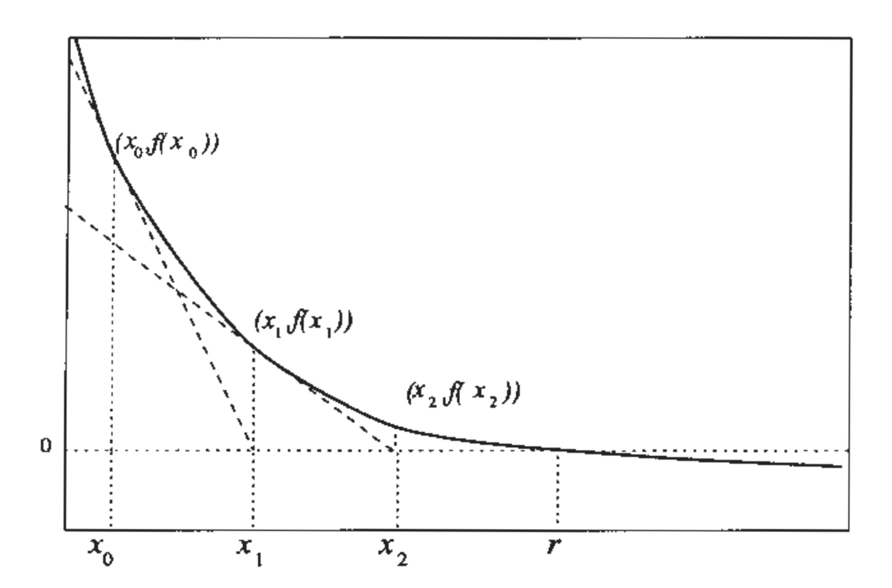


Figure 2.4: Interprétation géométrique de la méthode de Newton

qui correspond au développement de Taylor de degré 1 autour de x_0. Cette droite coupe l'axe des x en $y = 0$, c'est-à-dire en:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

qui devient la nouvelle valeur estimée de la solution. On reprend ensuite le même raisonnement à partir du point $(x_1, f(x_1))$ et ainsi de suite.

2.4.2 Analyse de convergence

La méthode de Newton est un cas particulier de la méthode des points fixes où:

$$g(x) = x - \frac{f(x)}{f'(x)}$$

Il n'est donc pas nécessaire de reprendre l'analyse de convergence à zéro. En effet, on sait que la convergence dépend de $g'(r)$ et on a dans ce cas précis:

$$g'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2} \tag{2.14}$$

Puisque $f(r) = 0$, r étant une racine, on a immédiatement $g'(r) = 0$ et donc une convergence au moins quadratique en vertu de la relation 2.9.

Remarque 2.10
Il faut noter que dans le cas où $f'(r)$ est également nul, le résultat précédent n'est plus vrai dans la mesure où $g'(r)$ pourra être différent de 0. Nous étudierons cette question en détail un peu plus loin. □

Pour s'assurer que la convergence de la méthode de Newton est bel et bien quadratique en général, il suffit de calculer $g''(r)$. On a, en vertu de l'équation 2.14:

$$g''(x) = \frac{[f'(x)f''(x) + f(x)f'''(x)](f'(x))^2 - 2f(x)f'(x)(f''(x))^2}{(f'(x))^4} \qquad (2.15)$$

On en conclut que puisque $f(r) = 0$:

$$g''(r) = \frac{f''(r)}{f'(r)}$$

et que $g''(r)$ n'a *a priori* aucune raison d'être nul. Il reste que l'on a supposé que $f'(r) \neq 0$, ce qui n'est pas toujours vrai. Enfin, de la relation 2.9 on déduit:

$$e_{n+1} \simeq \frac{g''(r)}{2}e_n^2 = \frac{f''(r)}{2f'(r)}e_n^2 \qquad (2.16)$$

qui démontre bien la convergence quadratique (si $f'(r) \neq 0$).

Exemple 2.11
Reprenons l'exemple où on doit calculer $\sqrt{2}$ en résolvant:

$$f(x) = x^2 - 2 = 0$$

Dans ce cas, $f'(x) = 2x$ et $f'(\sqrt{2}) = 2\sqrt{2} \neq 0$. On doit donc s'attendre à une convergence quadratique. Voici les résultats obtenus à l'aide de la méthode de Newton.

n	x_n	$\|e_n\|$	$\|\frac{e_n}{e_{n-1}}\|$	$\|\frac{e_n}{e_{n-1}^2}\|$
0	2,000 0000	$0{,}5858 \times 10^{+0}$	—	—
1	1,500 0000	$0{,}8578 \times 10^{-1}$	$0{,}1464 \times 10^{+0}$	0,0868
2	1,416 6666	$0{,}2453 \times 10^{-2}$	$0{,}2860 \times 10^{-1}$	0,3333
3	1,414 2157	$0{,}2124 \times 10^{-5}$	$0{,}8658 \times 10^{-3}$	0,3529
4	1,414 2136	$0{,}1594 \times 10^{-11}$	$0{,}7508 \times 10^{-6}$	0,3535

Le tableau précédent est encore une fois très instructif. On remarque que e_n tend vers 0 et que le ratio $|e_n/e_{n-1}|$ tend aussi vers 0 (c'est-à-dire vers $g'(r)$, qui est 0 dans ce cas). De plus, le ratio $|e_n/e_{n-1}^2|$ tend vers à peu près 0,3535. Encore une fois, ce nombre n'est pas arbitraire et correspond à:

$$\frac{f''(r)}{2f'(r)} = \frac{f''(\sqrt{2})}{2f'(\sqrt{2})} = \frac{1}{2\sqrt{2}} \simeq 0{,}353\,553$$

en vertu de la relation 2.16.

●●●●●

Remarque 2.11
Tout comme c'était le cas avec la méthode des points fixes, la convergence de la méthode de Newton dépend de la valeur initiale x_0. Malgré ses belles propriétés de convergence, une mauvaise valeur initiale peut provoquer la divergence de cette méthode. □

2.4.3 Cas des racines multiples

Il arrive parfois que la méthode de Newton ne converge pas aussi vite que l'on s'y attendait. Cela est souvent le signe d'une *racine multiple*, dont nous rappelons la définition.

Définition 2.7
Une racine r de la fonction $f(x)$ est dite *de multiplicité* m si la fonction $f(x)$ peut s'écrire sous la forme:

$$f(x) = (x - r)^m h(x) \tag{2.17}$$

et ce pour une fonction $h(x)$ qui vérifie $h(r) \neq 0$.

Si on a une racine de multiplicité m en $x = r$, on peut mettre en facteur un terme de la forme $(x - r)^m$ de telle sorte que le reste $h(x)$ ne s'annule pas en $x = r$.

Il est facile de démontrer, en utilisant un développement de Taylor autour de r, que le résultat suivant est vrai.

Théorème 2.2

Une racine r est de multiplicité m (où m est un entier) si et seulement si:

$$f(r) = f'(r) = f''(r) = \cdots = f^{(m-1)}(r) = 0, \quad f^{(m)}(r) \neq 0 \tag{2.18}$$

c'est-à-dire si la fonction de même que ses $(m-1)$ premières dérivées s'annulent en r (la dérivée d'ordre m ne doit pas s'annuler en r). □

Exemple 2.12

La fonction $f(x) = x \sin x$ possède une racine de multiplicité 2 en $x = 0$. En effet:

$$\begin{aligned} f(x) &= x \sin x \\ f'(x) &= \sin x + x \cos x \\ f''(x) &= 2 \cos x - x \sin x \end{aligned}$$

et on conclut aisément que $f(0) = 0$, $f'(0) = 0$ et $f''(0) \neq 0$.

● ● ● ● ●

Qu'arrive-t-il si on applique la méthode de Newton à ce cas? Rappelons que:

$$g'(x) = \frac{f(x) f''(x)}{(f'(x))^2}$$

et que, si on est en présence d'une racine de multiplicité m, on a:

$$\begin{aligned} f(x) &= (x-r)^m h(x) \\ f'(x) &= m(x-r)^{m-1} h(x) + (x-r)^m h'(x) \\ f''(x) &= m(m-1)(x-r)^{m-2} h(x) \\ &\quad + 2m(x-r)^{m-1} h'(x) + (x-r)^m h''(x) \end{aligned}$$

En remplaçant et en simplifiant le facteur $(x-r)^{2m-2}$, on trouve:

$$g'(x) = \frac{h(x)[m(m-1)h(x) + 2m(x-r)h'(x) + (x-r)^2 h''(x)]}{[mh(x) + (x-r)h'(x)]^2}$$

Cela entraîne que:

$$g'(r) = \frac{h(r)[m(m-1)h(r) + 0]}{m^2 (h(r))^2}$$

Puisque $h(r) \neq 0$, on peut simplifier cette relation et obtenir:

$$g'(r) = 1 - \frac{1}{m}$$

On constate maintenant que $g'(r) = 0$ seulement si $m = 1$, c'est-à-dire si on a une racine simple (de multiplicité 1). La convergence ne sera donc quadratique que pour les racines simples. Si $m \neq 1$, la méthode de Newton converge linéairement avec un taux de convergence de $1 - 1/m$. On remarque aussi que plus m est grand, plus la convergence est lente, car $g'(r)$ est de plus en plus près de 1.

Exemple 2.13
On considère la résolution de

$$f(x) = x^3 - 5x^2 + 7x - 3 = 0$$

En partant de $x_0 = 0$, on obtient le tableau suivant.

n	x_n	$\lvert e_n \rvert$	$\lvert \frac{e_n}{e_{n-1}} \rvert$	$\lvert \frac{e_n}{e_{n-1}^2} \rvert$
0	0,000 0000	1,0000	—	—
1	0,428 5714	0,5714	0,5714	0,5714
2	0,685 7143	0,3142	0,5499	0,9625
3	0,832 8654	0,1671	0,5318	1,6926
4	0,913 3299	0,0866	0,5185	3,1017
5	0,955 7833	0,0442	0,5102	5,8864
6	0,977 6551	0,0223	0,5045	11,429

On voit tout de suite que la convergence vers la racine $r = 1$ est lente. On vérifie aisément que $f(1) = f'(1) = 0$ et donc que 1 est une racine de multiplicité 2 ($m = 2$). Cela est confirmé par la quatrième colonne du tableau, qui doit normalement converger vers $1 - 1/m$, c'est-à-dire vers 0,5 dans ce cas précis. On note enfin que les valeurs de la dernière colonne semblent augmenter sans cesse et tendre vers l'infini. Cela indique une fois de plus que la convergence est linéaire et non quadratique.

● ● ● ● ●

Il existe des moyens de récupérer la convergence quadratique dans le cas de racines multiples. Il suffit en effet de transformer le problème en un

problème équivalent ayant les mêmes racines, mais de multiplicité 1. Dans cette optique, considérons la fonction:

$$u(x) = \frac{f(x)}{f'(x)}$$

On a immédiatement:

$$u(x) = \frac{(x-r)^m h(x)}{m(x-r)^{m-1}h(x) + (x-r)^m h'(x)} = \frac{(x-r)h(x)}{mh(x) + (x-r)h'(x)}$$

et

$$\begin{aligned} u'(x) &= \frac{[h(x) + (x-r)h'(x)][mh(x) + (x-r)h'(x)]}{[mh(x) + (x-r)h'(x)]^2} \\ &\quad - \frac{[(x-r)h(x)][mh'(x) + h'(x) + (x-r)h''(x)]}{[mh(x) + (x-r)h'(x)]^2} \end{aligned}$$

Puisque $h(r) \neq 0$, on a $u(r) = 0$ mais aussi $u'(r) = 1/m \neq 0$. r est donc une racine simple de $u(x)$, mais une racine multiple de $f(x)$. On peut dès lors appliquer l'algorithme de Newton à la fonction $u(x)$ pour trouver cette racine. L'algorithme devient:

$$x_{n+1} = x_n - \frac{u(x_n)}{u'(x_n)} = x_n - \frac{\frac{f(x_n)}{f'(x_n)}}{\frac{(f'(x_n))^2 - f(x_n)f''x_n)}{(f'(x_n))^2}}$$

ou plus succinctement:

$$x_{n+1} = x_n - \frac{f(x_n)f'(x_n)}{(f'(x_n))^2 - f(x_n)f''(x_n)} \tag{2.19}$$

On note que cet algorithme requiert la connaissance de $f(x)$, de $f'(x)$ et de $f''(x)$, ce qui peut rendre laborieux le processus de résolution. Si on reprend le problème de l'exemple précédent en utilisant cette fois l'algorithme 2.19, on retrouve la convergence quadratique, comme en témoignent les résultats suivants.

n	x_n	$\lvert e_n \rvert$	$\lvert \frac{e_n}{e_{n-1}} \rvert$	$\lvert \frac{e_n}{e_{n-1}^2} \rvert$
0	0,000 000	$0{,}1000 \times 10^{+1}$	—	—
1	1,105 263	$0{,}1053 \times 10^{+0}$	$0{,}1052 \times 10^{+0}$	0,1053
2	1,003 082	$0{,}3081 \times 10^{-2}$	$0{,}2927 \times 10^{-1}$	0,2781
3	1,000 002	$0{,}2382 \times 10^{-5}$	$0{,}7729 \times 10^{-3}$	0,2508

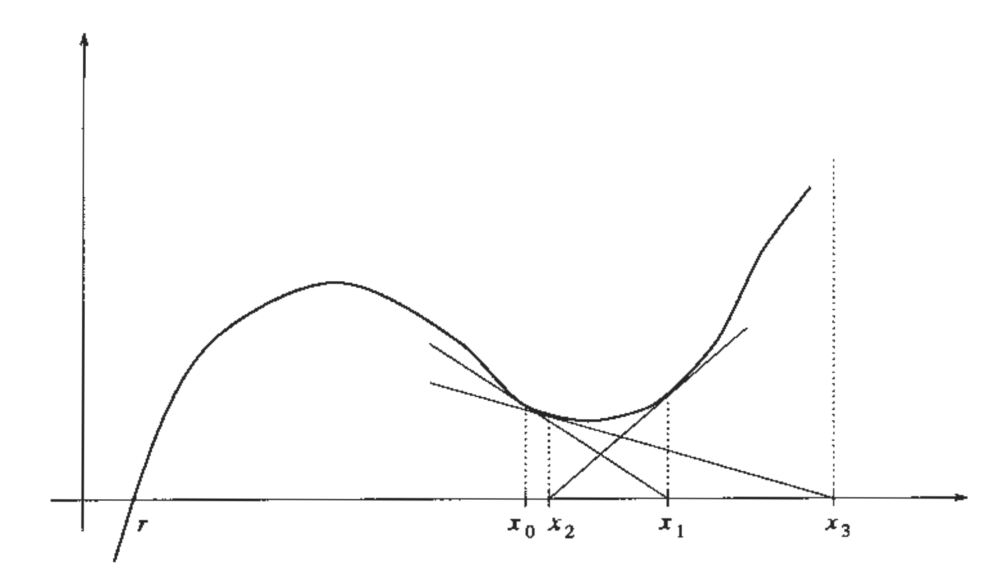


Figure 2.5: Cas pathologique pour la méthode de Newton

Remarque 2.12

Il existe un autre algorithme qui permet de récupérer la convergence quadratique de la méthode de Newton, mais il exige de connaître à l'avance la multiplicité m de la racine recherchée. Cela est évidemment très rare en pratique. On retrouvera cet algorithme dans les exercices de fin de chapitre. □

Exemple 2.14

Les racines multiples ne sont pas la seule source de difficultés que l'on peut rencontrer avec la méthode de Newton. Quelques cas pathologiques, comme celui qu'illustre la figure 2.5, aboutissent à la divergence de l'algorithme. Le choix de la valeur initiale x_0 est primordial, car la convergence de l'algorithme en dépend fortement. Dans l'exemple de la figure 2.5, une valeur de x_0 plus près de la racine r permettrait de retrouver la convergence.

♦ ♦ ♦ ♦ ♦

2.5 Méthode de la sécante

La méthode de Newton possède de grands avantages, mais elle nécessite le calcul de la dérivée de $f(x)$. Si la fonction $f(x)$ est complexe, cette dérivée peut être difficile à évaluer et peut résulter en une expression complexe. On contourne cette difficulté en remplaçant le calcul de la pente $f'(x_n)$ de la droite tangente à la courbe par l'expression suivante:

$$f'(x_n) \simeq \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

Cela revient à utiliser la droite sécante passant par les points $(x_n, f(x_n))$ et $(x_{n-1}, f(x_{n-1}))$ au lieu de la droite tangente passant par $(x_n, f(x_n))$. Ce choix est représenté à la figure 2.6. Il en résulte l'algorithme suivant.

Algorithme 2.5: Algorithme de la méthode de la sécante

1. Étant donné ϵ, un critère d'arrêt
2. Étant donné N, le nombre maximal d'itérations
3. Étant donné x_0 et x_1, deux valeurs initiales de la solution
4. Effectuer:
$$x_{n+1} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{(f(x_n) - f(x_{n-1}))}$$
5. Si $\frac{|x_{n+1} - x_n|}{|x_{n+1}|} < \epsilon$:
   - convergence atteinte
   - écrire la solution x_{n+1}
   - arrêt
6. Si le nombre maximal d'itérations N est atteint:
   - convergence non atteinte en N itérations
   - arrêt
7. retour à l'étape 4 □

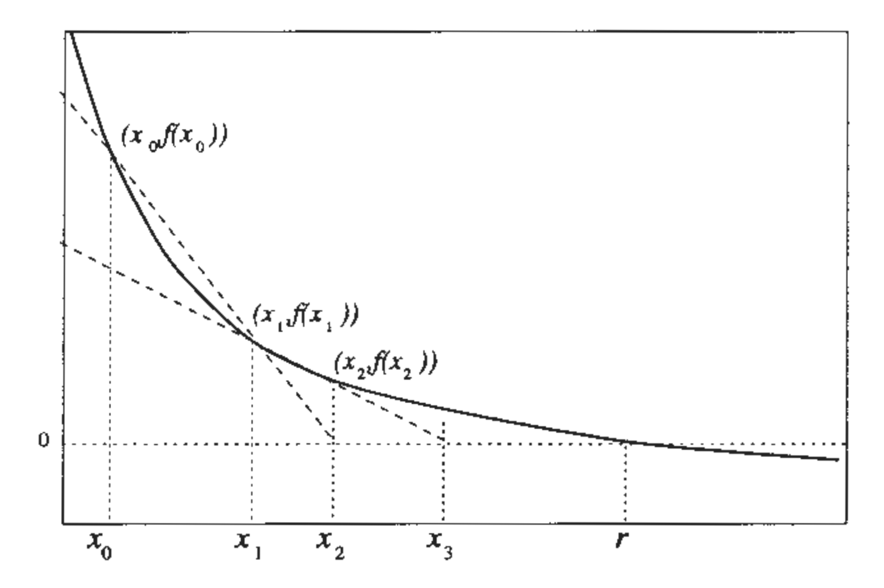


Figure 2.6: Interprétation géométrique de la méthode de la sécante

Remarque 2.13
Plusieurs remarques s'imposent au sujet de cet algorithme.

1. La dérivée de $f(x)$ n'apparaît plus dans l'algorithme.

2. Il faut fournir au départ 2 valeurs initiales. C'est ce qu'on appelle un *algorithme à deux pas*.

3. On choisit les valeurs initiales le plus près possible de la racine recherchée. *Il n'est cependant pas nécessaire qu'il y ait un changement de signe dans l'intervalle* $[x_0, x_1]$, comme c'est le cas avec la méthode de la bissection.

4. L'analyse de convergence de cet algorithme est plus délicate que celle de la méthode de Newton. On est cependant en mesure d'avancer que *la convergence quadratique est perdue, mais que la convergence est plus que linéaire*. En effet, on peut montrer (voir Conte et de Boor, réf. [6]) que:

   $$|e_{n+1}| \simeq C|e_n|^{\frac{1+\sqrt{5}}{2}} = C|e_n|^{1{,}618\,033\cdots}$$

   Cet ordre de convergence quelque peu étonnant vient justement du fait que la méthode est à deux pas. □

Illustrons cette méthode à l'aide d'un exemple.

Exemple 2.15
On cherche à résoudre:

$$e^{-x} - x = 0$$

que nous avons déjà abordé par d'autres méthodes. En prenant $x_0 = 0$ et $x_1 = 1$, on trouve à la première itération:

$$\begin{aligned} x_2 &= x_1 - \frac{(e^{-x_1} - x_1)(x_1 - x_0)}{(e^{-x_1} - x_1) - (e^{-x_0} - x_0)} \\ &= 1 - \frac{(e^{-1} - 1)(1 - 0)}{(e^{-1} - 1) - (e^0 - 0)} = 0{,}612\,6998 \end{aligned}$$

Les résultats sont compilés dans le tableau suivant.

n	x_n	$\|e_n\|$	$\left\|\frac{e_n}{e_{n-1}}\right\|$	$\left\|\frac{e_n}{e_{n-1}^{1,618}}\right\|$	$\left\|\frac{e_n}{e_{n-1}^2}\right\|$
0	0,000 0000	$0{,}5671 \times 10^{+0}$	—	—	—
1	1,000 0000	$0{,}4328 \times 10^{+0}$	$0{,}7632 \times 10^{+0}$	1,0835	1,342
2	0,612 6998	$0{,}4555 \times 10^{-1}$	$0{,}1052 \times 10^{+0}$	0,1766	0,243
3	0,563 8384	$0{,}3305 \times 10^{-2}$	$0{,}7254 \times 10^{-1}$	0,4894	1,592
4	0,567 1704	$0{,}2707 \times 10^{-4}$	$0{,}8190 \times 10^{-2}$	0,2796	2,478
5	0,567 1433	$0{,}1660 \times 10^{-7}$	$0{,}6134 \times 10^{-3}$	0,4078	22,66
6	0,567 1433	$\simeq 0$	$\simeq 0$		

La chose la plus importante à remarquer est que le ratio $|e_n/e_{n-1}|$ tend vers 0 mais que le ratio $|e_n/e_{n-1}^2|$ tend vers l'infini, ce qui confirme que l'ordre de convergence se trouve quelque part entre 1 et 2. On remarque que le quotient $|e_n/e_{n-1}^{1,618}|$ semble se stabiliser autour de 0,4 bien que la précision soit insuffisante pour être plus affirmatif. Il semble bien que cette suite ne tend ni vers 0 ni vers l'infini, ce qui confirme que l'ordre de convergence se situe autour de 1,618.

♦ ♦ ♦ ♦ ♦

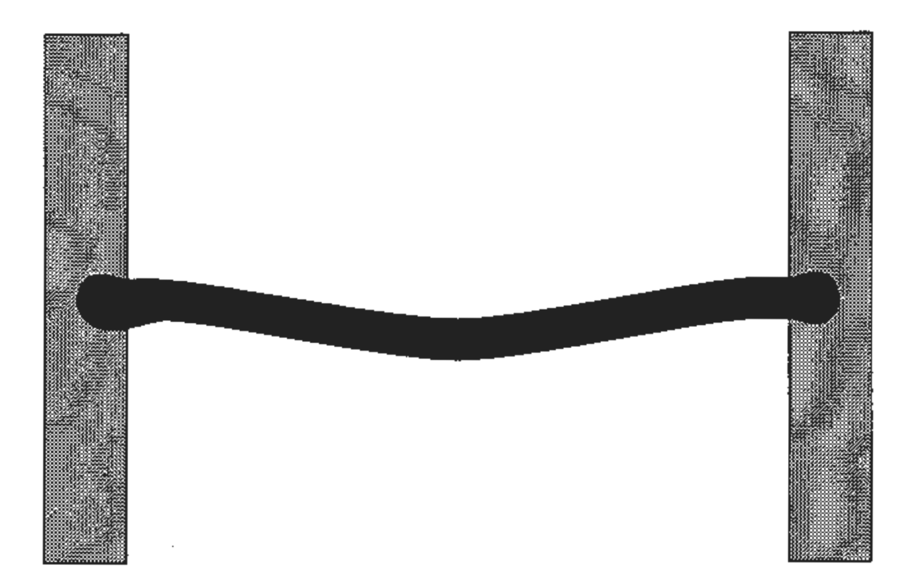


Figure 2.7: Problème de la poutre encastrée

2.6 Applications

Nous présentons dans cette section quelques exemples d'applications des méthodes numériques vues dans ce chapitre à des problèmes d'ingénierie. Chaque problème est brièvement décrit de manière à donner une idée assez précise du contexte, sans toutefois s'attarder sur les détails.

2.6.1 Modes de vibration d'une poutre

Une poutre de longueur L encastrée aux deux extrémités (voir la figure 2.7) subit une déformation au temps $t = 0$ et se met par la suite à vibrer. La déformation $u(x,t)$ de la poutre à la position x et au temps t est solution de:

$$\frac{\partial^2 u}{\partial t^2} + c^2 \frac{\partial^4 u}{\partial x^4} = 0 \tag{2.20}$$

qui est une équation aux dérivées partielles d'ordre 4. La constante c dépend de l'élasticité de la poutre. Les conditions aux limites traduisent l'état de la poutre à chaque extrémité. Nous avons choisi le cas où celle-ci est encastrée, ce qui impose les conditions:

$$\begin{array}{ll} u(0,t) = u(L,t) = 0 & \text{(fixée aux 2 extrémités)} \\ u_x(0,t) = u_x(L,t) = 0 & \text{(encastrée aux 2 extrémités)} \end{array} \tag{2.21}$$

Des conditions relatives à la déformation $u(x,0)$ et à la vitesse $u_t(x,0)$ initiales complètent ce système.

Une méthode classique de résolution de l'équation 2.20 consiste à séparer les variables (voir Kreyszig, réf. [15]) et à rechercher des solutions de la forme:

$$u(x,t) = F(x)G(t) \tag{2.22}$$

Les conditions aux limites 2.21 imposent des conditions à la fonction $F(x)$, qui sont:

$$\begin{array}{c} F(0) = F(L) = 0 \\ F'(0) = F'(L) = 0 \end{array} \tag{2.23}$$

En remplaçant l'équation 2.22 dans l'équation 2.20, on obtient:

$$\frac{F''''(x)}{F(x)} = \frac{G''(t)}{c^2G(t)}$$

où on remarque l'égalité d'une fonction de x et d'une fonction de t, pour tout x et t. Cela n'est possible que si les deux fonctions sont égales à une constante. On peut vérifier que cette constante ne peut être négative ou nulle et nous la notons β^4. On en vient à considérer les deux équations différentielles ordinaires:

$$\begin{array}{c} F''''(x) - \beta^4 F(x) = 0 \\ G''(t) + c^2\beta^4 G(t) = 0 \end{array}$$

dont les solutions respectives sont de la forme[2]:

$$\begin{array}{l} F(x) = A\cos(\beta x) + B\sin(\beta x) + C\cosh(\beta x) + D\sinh(\beta x) \\ G(t) = a\cos(c\beta^2 t) + b\sin(c\beta^2 t) \end{array} \tag{2.24}$$

On conclut de plus que:

$$F'(x) = \beta(-A\sin(\beta x) + B\cos(\beta x) + C\sinh(\beta x) + D\cosh(\beta x))$$

[2] $\cosh x$ et $\sinh x$ sont les fonctions hyperboliques définies par:

$$\cosh x = \frac{e^x + e^{-x}}{2} \quad \text{et} \quad \sinh x = \frac{e^x - e^{-x}}{2}$$

d'où l'on tire les propriétés classiques (laissées en exercices):

$$(\cosh x)' = \sinh x, \qquad (\sinh x)' = \cosh x \quad \text{et} \quad \cosh^2 x - \sinh^2 x = 1$$

La constante β est pour le moment arbitraire. Les conditions 2.23 imposent les contraintes:

$$\begin{array}{ll} F(0) = A + C = 0 & \text{c.-à-d. } C = -A \\ F'(0) = \beta(B + D) = 0 & \text{c.-à-d. } D = -B \end{array}$$

La fonction $F(x)$ peut alors s'écrire:

$$F(x) = A(\cos(\beta x) - \cosh(\beta x)) + B(\sin(\beta x) - \sinh(\beta x))$$

et sa dérivée:

$$F'(x) = \beta(A(-\sin(\beta x) - \sinh(\beta x)) + B(\cos(\beta x) - \cosh(\beta x)))$$

Les deux dernières conditions aux limites imposent:

$$\begin{array}{rcl} F(L) & = & A(\cos(\beta L) - \cosh(\beta L)) + B(\sin(\beta L) - \sinh(\beta L)) = 0 \\ F'(L) & = & \beta(A(-\sin(\beta L) - \sinh(\beta L)) + B(\cos(\beta L) - \cosh(\beta L))) = 0 \end{array}$$

ce qui peut encore s'exprimer sous la forme du système linéaire:

$$\left[\begin{array}{cc} (\cos(\beta L) - \cosh(\beta L)) & (\sin(\beta L) - \sinh(\beta L)) \\ -\beta(\sin(\beta L) + \sinh(\beta L)) & \beta(\cos(\beta L) - \cosh(\beta L)) \end{array} \right] \left[\begin{array}{c} A \\ B \end{array} \right] = \left[\begin{array}{c} 0 \\ 0 \end{array} \right]$$

Si la matrice précédente est inversible, la seule solution possible est $A = B = 0$, ce qui signifie que $F(x) = 0$, qui est une solution triviale. Pour obtenir des solutions non triviales, le déterminant doit être nul, ce qui signifie que:

$$\beta(\cos(\beta L) - \cosh(\beta L))^2 + \beta(\sin(\beta L) - \sinh(\beta L))(\sin(\beta L) + \sinh(\beta L)) = 0$$

c'est-à-dire

$$1 - \cos(\beta L)\cosh(\beta L) = 0 \tag{2.25}$$

Les seules valeurs intéressantes de β sont celles qui vérifient cette équation non linéaire. On est amené à rechercher les racines de la fonction:

$$f(x) = 1 - \cos x \cosh x$$

Cette fonction varie fortement, car $\cosh x$ prend des valeurs très grandes tandis que $\cos x$ oscille du positif au négatif. Pour simplifier le traitement numérique, une bonne stratégie consiste à considérer la fonction:

$$f_1(x) = \frac{1 - \cos x \cosh x}{\cosh x} = \frac{1}{\cosh x} - \cos x$$

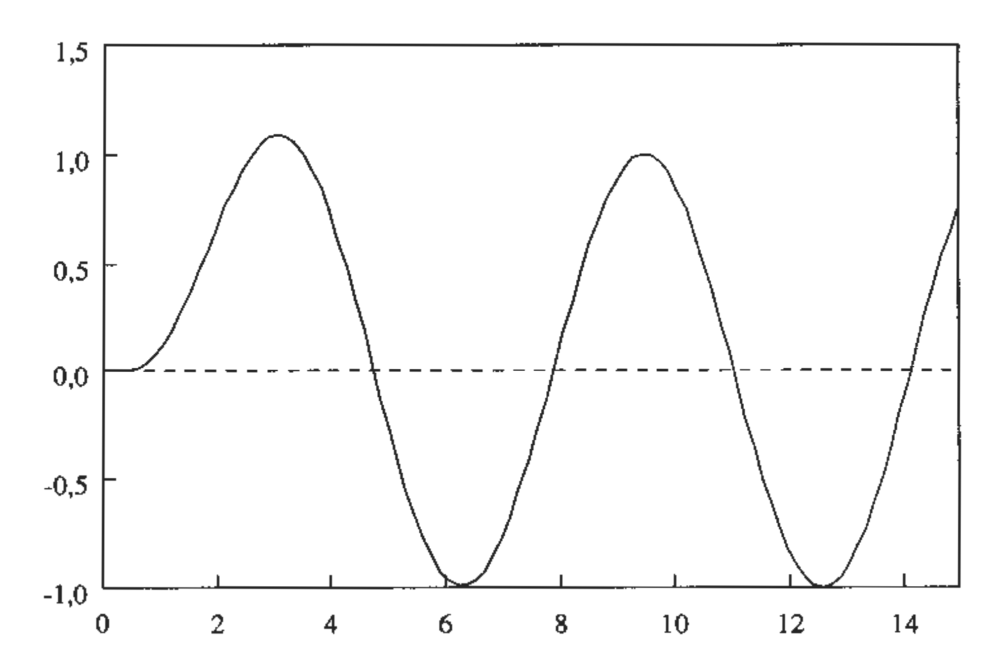


Figure 2.8: Fonction $f_1(x) = (1 - \cos x \cosh x)/\cosh x$

qui possède les mêmes racines que $f(x)$ et qui est illustrée à la figure 2.8. Heureusement, il est suffisant de trouver les premières racines seulement, qui correspondent aux modes de vibration les plus importants.

On constate aisément à l'aide de la figure qu'il y a un changement de signe dans les intervalles $[3, 5]$, $[6, 8]$ et $[10, 12]$. La méthode de la bissection appliquée à chacun de ces trois intervalles converge vers $x_1 = 4{,}730$, $x_2 = 7{,}853$ et $x_3 = 10{,}996$, correspondant aux trois premiers modes de vibration. On obtient les valeurs respectives de β en divisant les x_i par la longueur L de la poutre. D'autres méthodes de résolution d'équations non linéaires que nous avons vues dans ce chapitre auraient pu donner des résultats similaires.

2.6.2 Premier modèle de viscosité

Les polymères (plastiques) sont largement utilisés pour la production d'objets de toutes sortes, allant des simples jouets jusqu'à bon nombre de pièces d'automobile. La mise en forme de ces polymères requiert une étape de plastification où le polymère est fondu dans le but de lui donner sa forme finale, très souvent par moulage. Un des paramètres fondamentaux de cette

étape est la viscosité. Les rhéologues ont pour tâche de déterminer comment varie cette viscosité η en fonction du taux de cisaillement $\dot{\gamma}$. Des appareils nommés rhéomètres permettent de mesurer la viscosité pour différentes valeurs du taux de cisaillement. On obtient alors des résultats de la forme suivante.

Taux de cisaillement $\dot{\gamma}_i(s^{-1})$	Viscosité $\eta_i(Pa \cdot s)$
0,0137	3220,0
0,0274	2190,0
0,0434	1640,0
0,0866	1050,0
0,137	766,0
0,274	490,0
0,434	348,0
0,866	223,0
1,37	163,0
2,74	104,0
4,34	76,7
5,46	68,1
6,88	58,2

Ces valeurs caractérisent une solution de 2 % de polyisobutylène dans du primol 355 (voir Carreau, De Kee et Chhabra, réf. [3]). On cherche ensuite à modéliser cette variation selon une loi aussi simple que possible. Un modèle très populaire est la *loi puissance* de la forme:

$$\eta = \eta_0 \dot{\gamma}^{\beta-1} \tag{2.26}$$

où η_0 est la *consistance* et β est l'*indice de pseudoplasticité*. Ces deux derniers paramètres sont inconnus et doivent être déterminés à partir des données du tableau. On doit choisir ces paramètres de façon rendre compte le mieux possible des données. Un moyen courant d'y parvenir consiste à minimiser la fonction:

$$F(\eta_0, \beta) = \frac{1}{2} \sum_{i=1}^{npt} (\eta_0 \dot{\gamma}_i^{\beta-1} - \eta_i)^2$$

où npt est le nombre de mesures. C'est ce qu'on appelle une méthode de *moindres carrés* qui permet de minimiser la distance entre les points de mesure et la courbe représentée par la relation 2.26.

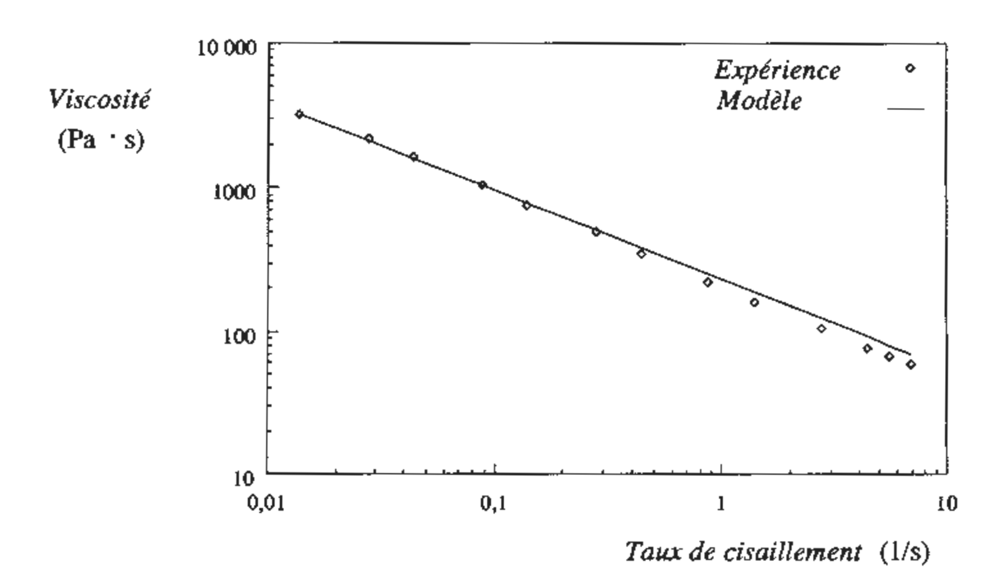


Figure 2.9: Loi puissance: $\eta = \eta_0 \dot{\gamma}^{\beta-1}$ ($\beta = 0{,}3797$, $\eta_0 = 228{,}34$)

L'écart minimal est atteint lorsque:

$$\frac{\partial F(\eta_0, \beta)}{\partial \eta_0} = \frac{\partial F(\eta_0, \beta)}{\partial \beta} = 0$$

On obtient ainsi les conditions d'optimalité[3]:

$$\frac{\partial F(\eta_0, \beta)}{\partial \eta_0} = \sum_{i=1}^{npt} (\eta_0 \dot{\gamma}_i^{\beta-1} - \eta_i) \dot{\gamma}_i^{\beta-1} = 0$$

$$\frac{\partial F(\eta_0, \beta)}{\partial \beta} = \sum_{i=1}^{npt} (\eta_0 \dot{\gamma}_i^{\beta-1} - \eta_i) \eta_0 \dot{\gamma}_i^{\beta-1} \ln \dot{\gamma}_i = 0$$

De la première équation, on tire une expression pour η_0 en fonction de β de la forme:

$$\eta_0 = \frac{\displaystyle\sum_{i=1}^{npt} \eta_i \dot{\gamma}_i^{\beta-1}}{\displaystyle\sum_{i=1}^{npt} \dot{\gamma}_i^{2\beta-2}} \tag{2.27}$$

[3]La dérivée par rapport à x de $a^{f(x)}$ est $a^{f(x)} f'(x) \ln a$.

Il reste donc à trouver β, solution de:

$$f(\beta) = \frac{\partial F(\eta_0, \beta)}{\partial \beta} = \sum_{i=1}^{npt} (\eta_0 \dot{\gamma}_i^{\beta-1} - \eta_i) \eta_0 \dot{\gamma}_i^{\beta-1} \ln \dot{\gamma}_i = 0$$

où η_0 est donné par l'équation 2.27. Il n'est pas facile d'établir la dérivée de la fonction $f(\beta)$. Dans le cas présent, la méthode de la sécante est presque aussi efficace que la méthode de Newton. L'indice de pseudoplasticité β est un nombre positif compris entre 0 et 1. À partir des valeurs initiales $\beta_0 = 0{,}5$ et $\beta_1 = 0{,}4$, la méthode de la sécante a convergé en 4 itérations vers $\beta = 0{,}3797$ ce qui donne une valeur $\eta_0 = 228{,}34$ en vertu de l'équation 2.27.

La figure 2.9 trace les points de mesure de même que la courbe de l'équation 2.26 pour ces valeurs. On remarque immédiatement que la correspondance n'est pas parfaite. Nous verrons au prochain chapitre un autre modèle qui colle davantage aux données rhéologiques, mais qui nécessite la résolution d'un système d'équations non linéaires.

2.7 Exercices

1. Faire trois itérations de la méthode de la bissection pour les fonctions suivantes et à partir des intervalles indiqués. Déterminer le nombre d'itérations nécessaires pour obtenir une solution dont le chiffre des millièmes est significatif.

   a) $f(x) = -0{,}9x^2 + 1{,}7x + 2{,}5$ dans l'intervalle $[2{,}8\,,3{,}0]$

   b) $f(x) = \dfrac{1 - 0{,}61x}{x}$ dans l'intervalle $[1{,}5\,,2{,}0]$

   c) $f(x) = x^2|\sin x| - 4{,}1$ dans l'intervalle $[0\,,4]$

   d) $f(x) = x^6 - x - 1$ dans l'intervalle $[1\,,2]$

2. Une variante de la méthode de la bissection, appelée *méthode de la fausse position*, consiste à remplacer le point milieu x_m de l'intervalle $[x_1, x_2]$ par le point d'intersection x_m^* de la droite joignant les points $(x_1, f(x_1))$ et $(x_2, f(x_2))$, avec l'axe des x. Illustrer à l'aide d'un graphique cette méthode. Obtenir l'équation de la droite et calculer son point d'intersection x_m^* avec l'axe des x. Modifier l'algorithme de la bissection en remplaçant x_m par x_m^*.

3. Reprendre l'exercice 1 en utilisant cette fois la méthode de la fausse position.

4. Obtenir la multiplicité m de la racine r des fonctions suivantes.

   a) $f(x) = x^2 - 2x + 1$, en $r = 1$

   b) $f(x) = x^3 - 2x^2 + x$, en $r = 0$

   c) $f(x) = x \sin x$, en $r = 0$

   d) $f(x) = \dfrac{\sin x}{x}$, en $r = 0$

5. Calculer les points fixes des fonctions suivantes et vérifier s'ils sont attractifs ou répulsifs.

   a) $g(x) = 4x - x^2$

   b) $g(x) = \sqrt{x}$

   c) $g(x) = \arcsin x$

   d) $g(x) = 5 + x - x^2$

6. Utiliser l'algorithme des points fixes avec les fonctions suivantes. Une fois la racine obtenue, calculer $|e_n|$ et $|e_n/e_{n-1}|$. Obtenir expérimentalement le taux de convergence de la méthode.

   a) $g(x) = 1 - x - \frac{x^2}{5}$ $(x_0 = 5)$

   b) $g(x) = \sqrt{1+x}$ $(x_0 = 1{,}5)$

7. Utiliser la méthode de Newton pour déterminer les racines des fonctions suivantes. Une fois la racine obtenue, calculer $|e_n|$, $|e_n/e_{n-1}|$ et $|e_n/e_{n-1}^2|$. Conclure sur la convergence de la méthode. Si la convergence est linéaire, modifier votre algorithme de façon à récupérer la convergence quadratique.

   a) $f(x) = x^3 - 2x^2 - 5$ $(x_0 = 3)$

   b) $f(x) = 0{,}51x - \sin x$ ($x_0 = 2$; et ensuite à partir de $x_0 = 1$)

   c) $f(x) = x^6 - x - 1$ $(x_0 = 1{,}5)$

   d) $f(x) = x^5 - 3x^4 + 4x^3 - 4x^2 + 3x - 1$ $(x_0 = 1{,}2)$

8. Faire 5 itérations de la méthode de la sécante pour les fonctions de l'exercice précédent.

9. Montrer que l'algorithme suivant permet de récupérer la convergence quadratique lorsque la multiplicité m de la racine est connue.

$$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)}$$

   Vérifier en premier lieu si cet algorithme converge vers une racine de $f(x)$. Montrer ensuite que la convergence est forcément quadratique.

10. On cherche à résoudre l'équation:

$$e^x - 3x^2 = 0$$

   qui possède les deux racines $r_1 = -0{,}458\,9623$ et $r_2 = 0{,}91$ ainsi qu'une troisième racine située près de 4. On vous propose les méthodes de

points fixes suivantes pour obtenir r_1.

$$1) \quad x = g_1(x) = -\sqrt{\frac{e^x}{3}}$$

$$2) \quad x = g_2(x) = -\left(\frac{e^x - 3x^2 - 3{,}385\,712\,869x}{3{,}385\,712\,869}\right)$$

$$3) \quad x = g_3(x) = -\left(\frac{e^x - 3x^2 - 3{,}761\,89x}{3{,}761\,89}\right)$$

a) Lesquelles, parmi ces trois méthodes de points fixes, sont susceptibles de converger vers r_1? (Ne pas faire les itérations.)

b) Déterminer celle qui produit une convergence quadratique vers r_1.

c) La méthode de la bissection convergera-t-elle vers l'une des racines si on prend [-1,0] comme intervalle de départ?

d) Utiliser la méthode de Newton pour déterminer la troisième racine avec 4 chiffres significatifs. Quel est l'ordre de convergence de cette méthode?

11. Évaluer la quantité:

$$s = \sqrt[3]{3 + \sqrt[3]{3 + \sqrt[3]{3 + \cdots}}}$$

Suggestion: Mettre cette équation au cube et obtenir une équation de la forme $f(s) = 0$. Résoudre cette dernière à l'aide de la méthode de Newton à partir de $s_0 = 1$.

12. On cherche à résoudre l'équation:

$$x^2 - 2 = 0$$

(dont la solution est $\sqrt{2}$) au moyen de la méthode de points fixes:

$$x_{n+1} = g(x_n)$$

où

$$g(x_n) = x_n - \rho(x_n^2 - 2)$$

et ρ est une constante.

a) Pour quelles valeurs de ρ cette méthode de points fixes est-elle convergente à l'ordre 1 (au moins)?

b) Quel est l'ordre de convergence pour $\rho = \sqrt{2}/4$?

c) Quel est l'ordre de convergence si $\rho = 3\sqrt{2}$?

13. On a calculé une racine de:

$$f(x) = x^3 + 4x^2 - 10$$

en utilisant l'algorithme de points fixes:

$$x_{n+1} = \frac{1}{2}\sqrt{10 - x_n^3}$$

On a obtenu les résultats suivants.

n	x_n	$\lvert e_n \rvert$	$\lvert \frac{e_n}{e_{n-1}} \rvert$
1	1,500 00	0,134 77	—
2	1,286 95	0,078 28	0,580 84
3	1,402 54	0,037 31	0,476 62
4	1,345 46	0,019 77	0,529 88
5	1,375 17	0,009 94	0,502 78
6	1,360 09	0,005 14	0,517 10
7	1,367 85	0,002 62	0,509 72
8	1,363 89	0,001 34	0,511 45
9	1,365 92	0,000 69	0,514 92
⋮	⋮	⋮	⋮
17	1,365 23	0,000 00	—

On a ontenu les résultats des deux dernières colonnes en considérant que la valeur exacte de la racine est $r = 1{,}365\,23$.

a) Expliquer pourquoi la méthode itérative précédente a convergé.

b) Les valeurs de $\lvert \frac{e_n}{e_{n-1}} \rvert$ semblent converger vers 0,51. Expliquer ce résultat et donner la valeur exacte vers laquelle le quotient $\lvert \frac{e_n}{e_{n-1}} \rvert$ devrait converger.

c) Quel est l'ordre de la méthode utilisée?

14. Une variante de la méthode de Newton pour résoudre des équations de la forme $f(x) = 0$ résulte en l'algorithme suivant:

$$\begin{cases} x_0 & \text{donné} \\ x_{n+1} = & x_n - \dfrac{f(x_n)}{f'(x_0)} \end{cases}$$

Note: La valeur de la dérivée apparaissant au dénominateur est fixée à $f'(x_0)$ pour toutes les itérations. Ce n'est donc pas la méthode de Newton.

a) Donner une interprétation géométrique de cette méthode.

b) On aimerait se servir de cette méthode pour évaluer la racine $r = \sqrt{2}$ de l'équation:

$$x^2 - 2 = 0$$

Donner une condition nécessaire sur x_0 pour que la méthode converge vers $\sqrt{2}$.

Suggestion: Considérer cette variante de la méthode de Newton comme une méthode de points fixes.

Chapitre 3

Systèmes d'équations algébriques

3.1 Introduction

Les systèmes d'équations algébriques jouent un rôle très important en ingénierie. On peut classer ces systèmes en deux grandes familles: les systèmes *linéaires* et les systèmes *non linéaires*.

Ici encore, les progrès de l'informatique et de l'analyse numérique permettent d'aborder des problèmes de taille prodigieuse. On résout couramment aujourd'hui des systèmes de plusieurs centaines de milliers d'inconnues. On rencontre ces applications en mécanique des fluides et dans l'analyse de structures complexes. On peut par exemple calculer l'écoulement de l'air autour d'un avion ou l'écoulement de l'eau dans une turbine hydraulique complète. On peut également analyser la résistance de la carlingue d'un avion à différentes contraintes extérieures et en vérifier numériquement la solidité.

Ces calculs complexes requièrent des méthodes sophistiquées comme les méthodes d'éléments finis (voir Reddy, réf. [24]). On obtient généralement des systèmes d'équations non linéaires de taille considérable, qu'on doit résoudre à l'aide de méthodes efficaces qui minimisent le temps de calcul et l'espace-mémoire requis.

Dans ce chapitre, nous allons aborder les principales méthodes de résolution des systèmes linéaires, à savoir la méthode d'élimination de Gauss et la décomposition LU. L'effet des erreurs dues à l'arithmétique flottante sera également étudié, et nous introduirons le concept de *conditionnement* d'une matrice.

Par la suite, nous verrons comment résoudre les systèmes non linéaires au moyen d'une suite de systèmes linéaires. C'est ce que nous appelons la linéarisation du problème.

3.2 Systèmes linéaires

De façon générale, la résolution d'un système d'équations linéaires consiste à trouver un vecteur $\vec{x} = [x_1 \ x_2 \ x_3 \ \cdots \ x_n]^T$ (*$\vec{x}$ dénotera toujours un vecteur colonne et l'indice supérieur T désignera sa transposée*) solution de:

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n & = & b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots + a_{3n}x_n & = & b_3 \\ \vdots & = & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n & = & b_n \end{array} \tag{3.1}$$

On peut utiliser la notation matricielle, qui est beaucoup plus pratique et surtout plus compacte. On écrit alors le système précédent sous la forme:

$$A\vec{x} = \vec{b} \tag{3.2}$$

où A est la matrice:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix}$$

et où $\vec{b} = [b_1 \ \ b_2 \ \ b_3 \ \cdots \ b_n]^T$ est le *membre de droite*. Bien entendu, la matrice A et le vecteur $\vec{b}$ sont connus. Il reste à déterminer le vecteur $\vec{x}$. Le problème 3.1 (ou 3.2) est un système de n équations et n inconnues. En pratique, la valeur de n varie considérablement et peut s'élever jusqu'à plusieurs centaines de milliers. Dans ce chapitre, nous nous limitons à des systèmes de petite taille, mais les stratégies développées sont valides quelle que soit la taille du système. Notons finalement que le coût de la résolution croît rapidement avec n.

Remarque 3.1
Dans la plupart des cas, nous traitons des *matrices non singulières* ou *inversibles*, c'est-à-dire dont la matrice inverse existe. Nous ne faisons pas non

plus de révision systématique de l'algèbre linéaire élémentaire que nous supposons connue. Ainsi, la solution de l'équation 3.2 peut s'écrire:

$$\vec{x} = A^{-1}\vec{b}$$

et la discussion peut sembler close. Nous verrons cependant que le calcul de la matrice inverse A^{-1} est plus difficile et plus long que la résolution du système linéaire de départ. □

Exemple 3.1
Considérons le système linéaire suivant:

$$\begin{aligned} 2x_1 + 3x_2 &= 8 \\ 3x_1 + 4x_2 &= 11 \end{aligned}$$

Pour le résoudre, on peut utiliser la méthode classique qui consiste à éliminer les équations une à une par ***substitution successive***. Dans un premier temps, on isole x_1 de la première équation:

$$x_1 = \frac{8 - 3x_2}{2}$$

que l'on substitue dans la deuxième équation:

$$3\left(\frac{8 - 3x_2}{2}\right) + 4x_2 = 11$$

ou encore

$$12 - 9x_2/2 + 4x_2 = 12 - 0{,}5x_2 = 11$$

On déduit alors facilement que $x_2 = 2$ et par la suite que $x_1 = 1$.

♦ ♦ ♦ ♦ ♦

Il est théoriquement possible d'étendre la substitution successive à des systèmes de grande taille. Il est cependant difficile de transcrire cette façon de faire sous forme d'algorithme (qui peut par la suite être programmé dans un langage informatique quelconque). Il est donc préférable de recourir à d'autres méthodes pour simplifier le système d'équations.

On peut d'abord se demander quels types de systèmes linéaires sont faciles à résoudre, et ce même s'ils sont de grande taille. Le cas le plus simple est sans doute celui des ***systèmes diagonaux***, c'est-à-dire dont la matrice A n'a de coefficients non nuls que sur la diagonale.

Exemple 3.2

Le système suivant:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 9 \end{bmatrix}$$

est très facile à résoudre. Il suffit de considérer séparément chaque ligne. On obtient ainsi $\vec{x} = [2 \;\; 1 \;\; 3]^T$. On voit tout de suite comment résoudre le cas général:

$$x_i = \frac{b_i}{a_{ii}} \quad \text{pour } i = 1,\ 2, \cdots, n$$

On remarque de plus que le système a une solution unique seulement si tous les termes diagonaux sont non nuls. Hélas, on rencontre rarement des systèmes diagonaux en pratique et il faudra travailler un peu plus pour s'attaquer aux applications.

◆◆◆◆◆

Le deuxième type de système simple est le ***système triangulaire inférieur*** ou ***supérieur***.

Définition 3.1
Une matrice est dite *triangulaire inférieure* (ou *supérieure*) si tous les a_{ij} (ou tous les a_{ji}) sont nuls pour $i < j$. Une matrice triangulaire inférieure a la forme type:

$$\begin{bmatrix} a_{11} & 0 & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & 0 & \cdots & 0 \\ a_{31} & a_{32} & a_{33} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n-1\,1} & a_{n-1\,2} & a_{n-1\,3} & \cdots & a_{n-1\,n} & 0 \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{n\,n-1} & a_{nn} \end{bmatrix}$$

Une matrice triangulaire supérieure est tout simplement la transposée d'une matrice triangulaire inférieure.

Les systèmes triangulaires sont également faciles à résoudre. Il suffit en effet de commencer par l'équation qui se trouve à la pointe du triangle (la première pour une matrice triangulaire inférieure et la dernière pour une matrice triangulaire supérieure) et de résoudre une à une les équations. On parle de *descente triangulaire* ou de *remontée triangulaire*, selon le cas.

Exemple 3.3
La descente triangulaire du système:

$$\begin{bmatrix} 3 & 0 & 0 \\ 1 & 2 & 0 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 7 \\ 14 \end{bmatrix}$$

consiste à résoudre la première équation:

$$x_1 = \frac{b_1}{a_{11}} = \frac{9}{3} = 3$$

Puisque x_1 est maintenant connue, on peut déterminer x_2:

$$x_2 = \frac{b_2 - a_{21}x_1}{a_{22}} = \frac{7 - (1)(3)}{2} = 2$$

La dernière équation s'écrit:

$$x_3 = \frac{b_3 - a_{31}x_1 - a_{32}x_2}{a_{33}} = \frac{14 - (3)(3) - (2)(2)}{1} = 1$$

● ● ● ● ●

De l'exemple précédent, on peut rapidement déduire le cas général pour la descente triangulaire:

$$\begin{aligned} x_1 &= b_1/a_{11} \\ x_i &= \frac{\left(b_i - \displaystyle\sum_{k=1}^{i-1} a_{ik}x_k\right)}{a_{ii}} \quad \text{pour } i = 2, 3, \cdots, n \end{aligned} \tag{3.3}$$

Pour la remontée triangulaire, on a:

$$\begin{aligned} x_n &= b_n/a_{nn} \\ x_i &= \frac{\left(b_i - \displaystyle\sum_{k=i+1}^{n} a_{ik}x_k\right)}{a_{ii}} \quad \text{pour } i = n-1, n-2, \cdots, 2, 1 \end{aligned} \tag{3.4}$$

Remarque 3.2
Les équations 3.3 et 3.4 sont valides si les a_{ii} sont tous non nuls. Dans le cas contraire, la matrice A n'est pas inversible et, donc, le système $A\vec{x} = \vec{b}$ n'a pas une solution unique. On se souvient en effet que le déterminant d'une matrice triangulaire inférieure (ou supérieure) est:

$$\text{dét } A_{triangulaire} = \prod_{i=1}^{n} a_{ii} \tag{3.5}$$

En d'autres mots, le déterminant est le produit des termes de la diagonale de A. Le produit est donc non nul seulement si aucun des a_{ii} n'est nul. □

Les matrices triangulaires sont primordiales pour la résolution des systèmes linéaires. Dans les sections qui suivent, nous voyons comment ramener un système linéaire quelconque à un ou plusieurs systèmes triangulaires. Nous abordons essentiellement deux méthodes dites *directes* au sens de la définition suivante.

Définition 3.2
Une méthode de résolution d'un système linéaire est dite *directe* si la solution du système peut être obtenue par cette méthode en un nombre fini et prédéterminé d'opérations.

Autrement dit, les méthodes directes permettent d'obtenir le résultat après un nombre connu de multiplications, divisions, additions et soustractions. On peut alors en déduire le temps de calcul nécessaire à la résolution (qui peut être très long si n est grand). Les méthodes directes s'opposent sur ce point aux méthodes dites *itératives*, qui peuvent converger en quelques itérations, converger en un très grand nombre d'itérations ou même diverger, selon le cas. Nous présentons quelques exemples de méthodes itératives à la fin du chapitre 4.

Les deux principales méthodes directes sont la *méthode d'élimination de Gauss* et la *décomposition LU*. Il s'agit en fait d'une seule et même méthode puisque la méthode d'élimination de Gauss est un cas particulier de décomposition LU. La stratégie de résolution est basée sur la question suivante: *Quelles opérations sont permises sur les lignes du système 3.1 pour le ramener à un système triangulaire?* Ou encore: Pour ramener un système linéaire quelconque à un système triangulaire, quels sont les coups permis, c'est-à-dire *ceux qui ne changent pas la solution du système de départ?* C'est à ces questions que nous répondons dans la section suivante.

3.3 Opérations élémentaires sur les lignes

Revenons au système:

$$A\vec{x} = \vec{b} \tag{3.6}$$

et voyons comment on peut le transformer sans en modifier la solution. La réponse est toute simple. On peut toujours multiplier (à gauche de chaque côté) les termes de cette relation par une matrice W *inversible*; la solution n'est pas modifiée puisque l'on peut remultiplier par W^{-1} pour revenir au système de départ. Ainsi:

$$WA\vec{x} = W\vec{b}$$

possède la même solution que le système 3.6.

Remarque 3.3
Ce résultat n'est plus vrai si la matrice W n'est pas inversible. On ne peut plus en effet revenir en arrière si la matrice W^{-1} n'existe pas. □

Exemple 3.4
Nous avons vu que la solution du système:

$$\begin{bmatrix} 3 & 0 & 0 \\ 1 & 2 & 0 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 7 \\ 14 \end{bmatrix}$$

est $\vec{x} = [3 \;\; 2 \;\; 1]^T$. Si on multiplie ce système par la matrice inversible:

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 2 & 3 \end{bmatrix}$$

on obtient le nouveau système:

$$\begin{bmatrix} 3 & 0 & 0 \\ 5 & 4 & 0 \\ 14 & 10 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 23 \\ 65 \end{bmatrix}$$

dont la solution est toujours $\vec{x} = [3 \;\; 2 \;\; 1]^T$. Par contre, si on multiplie le système de départ par la matrice non inversible:

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 2 & 0 \end{bmatrix}$$

on obtient le système singulier:

$$\begin{bmatrix} 3 & 0 & 0 \\ 5 & 4 & 0 \\ 5 & 4 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 23 \\ 23 \end{bmatrix}$$

qui possède une infinité de solutions, la dernière équation étant redondante.

● ● ● ● ●

Pour transformer un système quelconque en système triangulaire, il suffit d'utiliser trois opérations élémentaires sur les lignes de la matrice. Ces trois opérations élémentaires correspondent à trois types de matrices W différents. C'est la base de la méthode d'élimination de Gauss.

L'approche suivie est similaire à celle de Burden et Faires (réf. [2]). On note $\vec{l}_i$, la ligne i de la matrice A. Cette notation est quelque peu ambiguë, car on se trouve de ce fait à placer une ligne de la matrice A dans un vecteur colonne. Cela n'empêche cependant pas la compréhension de la suite.

Les trois opérations élémentaires dont on a besoin sont les suivantes:

1. Opération ($\vec{l}_i \leftarrow \lambda\vec{l}_i$): remplacer la ligne i par un multiple d'elle-même.
2. Opération ($\vec{l}_i \leftrightarrow \vec{l}_j$): intervertir la ligne i et la ligne j.
3. Opération ($\vec{l}_i \leftarrow \vec{l}_i + \lambda\vec{l}_j$): remplacer la ligne i par la ligne i plus un multiple de la ligne j.

Ces trois opérations élémentaires sont permises car elles équivalent à multiplier le système 3.6 par une matrice inversible.

3.3.1 Multiplication d'une ligne par un scalaire

Remplacer la ligne i par un multiple d'elle-même ($\vec{l}_i \leftarrow \lambda\vec{l}_i$) revient à multiplier le système linéaire 3.6 par une matrice diagonale inversible $W = M(\vec{l}_i \leftarrow \lambda\vec{l}_i)$, dont tous les éléments diagonaux sont 1, sauf a_{ii}, qui vaut λ. Tous les autres termes sont nuls. Cette matrice a pour effet de multiplier la ligne i par le scalaire λ.

Remarque 3.4
Le déterminant de la matrice diagonale $M(\vec{l}_i \leftarrow \lambda\vec{l}_i)$ est λ. La matrice est donc inversible si $\lambda \neq 0$. □

Remarque 3.5
La matrice inverse de $M(\vec{l}_i \leftarrow \lambda\vec{l}_i)$ est simplement $M(\vec{l}_i \leftarrow \lambda^{-1}\vec{l}_i)$, c'est-à-dire:

$$M^{-1}(\vec{l}_i \leftarrow \lambda\vec{l}_i) = M(\vec{l}_i \leftarrow (1/\lambda)\vec{l}_i) \tag{3.7}$$

Il suffit donc de remplacer λ par $1/\lambda$ pour inverser la matrice. □

Exemple 3.5
Soit le système:

$$\begin{bmatrix} 3 & 1 & 2 \\ 6 & 4 & 1 \\ 5 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 11 \\ 10 \end{bmatrix} \tag{3.8}$$

dont la solution est $\vec{x} = [1 \ 1 \ 1]^T$. Si on souhaite multiplier la ligne 2 par un facteur 3, cela revient à multiplier le système par la matrice:

$$M(\vec{l}_2 \leftarrow 3\vec{l}_2) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

et on obtient:

$$\begin{bmatrix} 3 & 1 & 2 \\ 18 & 12 & 3 \\ 5 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 33 \\ 10 \end{bmatrix}$$

La solution de ce nouveau système reste la même que celle du système de départ puisque la matrice $M(\vec{l}_2 \leftarrow 3\vec{l}_2)$ est inversible (et son déterminant est 3).

● ● ● ● ●

3.3.2 Permutation de deux lignes

L'opération élémentaire qui consiste à intervertir deux lignes ($\vec{l}_i \leftrightarrow \vec{l}_j$) est également connue sous le nom de *permutation de lignes.* Cette opération est équivalente à la multiplication du système 3.1 par une matrice inversible $W = P(\vec{l}_i \leftrightarrow \vec{l}_j)$, qui contient des 1 sur la diagonale, sauf à la ligne i, où le 1 est dans la colonne j, et à la ligne j, où le 1 est dans la colonne i. Tous les autres termes sont nuls.

Exemple 3.6
On veut intervertir la ligne 2 et la ligne 3 du système de l'exemple précédent. Il suffit de le multiplier par la matrice:

$$P(\vec{l}_2 \leftrightarrow \vec{l}_3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

et on obtient:

$$\begin{bmatrix} 3 & 1 & 2 \\ 5 & 4 & 1 \\ 6 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \\ 11 \end{bmatrix}$$

●●●●●

La matrice $P(\vec{l}_i \leftrightarrow \vec{l}_j)$ est inversible. Pour obtenir son inverse, il suffit de réfléchir une seconde. En effet, quelle est l'opération inverse de celle qui inverse deux lignes, sinon l'inversion des deux mêmes lignes?

Remarque 3.6
L'inverse de la matrice $P(\vec{l}_i \leftrightarrow \vec{l}_j)$ est donc la matrice $P(\vec{l}_i \leftrightarrow \vec{l}_j)$ elle-même, c'est-à-dire:

$$P^{-1}(\vec{l}_i \leftrightarrow \vec{l}_j) = P(\vec{l}_i \leftrightarrow \vec{l}_j) \quad \square \tag{3.9}$$

Remarque 3.7
On montre assez facilement que le déterminant de $P(\vec{l}_i \leftrightarrow \vec{l}_j)$ est -1. *Lorsque l'on permute deux lignes, le déterminant du système de départ change de signe.* □

3.3.3 Opération $(\vec{l}_i \leftarrow \vec{l}_i + \lambda\vec{l}_j)$

La dernière opération élémentaire consiste à remplacer la ligne i par la ligne i plus un multiple de la ligne j $(\vec{l}_i \leftarrow \vec{l}_i + \lambda\vec{l}_j)$. Cela est encore une fois équivalent à multiplier le système de départ par une matrice inversible $W = T(\vec{l}_i \leftarrow \vec{l}_i + \lambda\vec{l}_j)$ qui vaut 1 sur toute la diagonale et 0 partout ailleurs, sauf a_{ij}, qui vaut λ.

Exemple 3.7
Dans le système 3.8, on souhaite remplacer la deuxième ligne par la deuxième ligne ($i = 2$) moins deux fois ($\lambda = -2$) la première ligne ($j = 1$). Il suffit alors de multiplier le système par:

$$T(\vec{l}_2 \leftarrow \vec{l}_2 - 2\vec{l}_1) = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

ce qui donne:

$$\begin{bmatrix} 3 & 1 & 2 \\ 0 & 2 & -3 \\ 5 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -1 \\ 10 \end{bmatrix}$$

● ● ● ● ●

Remarque 3.8
La matrice $T(\vec{l}_i \leftarrow \vec{l}_i + \lambda\vec{l}_j)$ est inversible. Pour obtenir son inverse, il suffit de remplacer λ par $-\lambda$, c'est-à-dire:

$$T^{-1}(\vec{l}_i \leftarrow \vec{l}_i + \lambda\vec{l}_j) = T(\vec{l}_i \leftarrow \vec{l}_i - \lambda\vec{l}_j) \qquad (3.10)$$

Cela signifie que pour revenir en arrière il suffit de soustraire la ligne que l'on vient d'ajouter. □

Remarque 3.9
On peut montrer facilement que le déterminant de $T(\vec{l}_i \leftarrow \vec{l}_i + \lambda\vec{l}_j)$ est 1. □

Remarque 3.10
Dans cet exemple, en additionnant le bon multiple de la ligne 1 à la ligne 2, on a introduit un 0 à la position a_{21}. En remplaçant la ligne 3 par la ligne 3 moins (5/3) fois la ligne 1 (ou encore $\vec{l}_3 \leftarrow \vec{l}_3 - (5/3)\vec{l}_1$), on introduirait un terme 0 à la position a_{31}. On peut ainsi transformer un système linéaire quelconque en système triangulaire. C'est là la base sur laquelle repose la méthode d'élimination de Gauss. □

Remarque 3.11
Des trois opérations élémentaires, seule l'opération $(\vec{l_i} \leftarrow \vec{l_i} + \lambda \vec{l_j})$ n'a pas d'effet sur le déterminant. La permutation de deux lignes en change le signe, tandis que la multiplication d'une ligne par un scalaire multiplie le déterminant par ce même scalaire. □

3.4 Élimination de Gauss

Tous les outils sont en place pour la résolution d'un système linéaire. Il suffit maintenant d'utiliser systématiquement les opérations élémentaires pour introduire des zéros sous la diagonale de la matrice A et obtenir ainsi un système triangulaire supérieur.

La validité de la méthode d'élimination de Gauss repose sur le fait que les opérations élémentaires consistent à multiplier le système de départ par une matrice inversible.

Remarque 3.12
En pratique, on ne multiplie jamais les systèmes considérés par les différentes matrices W, car ce serait trop long. Il faut cependant garder en tête que les opérations effectuées sont équivalentes à cette multiplication. □

La méthode d'élimination de Gauss consiste à éliminer tous les termes sous la diagonale de la matrice A. Avant de considérer un exemple, introduisons la *matrice augmentée*.

Définition 3.3
La *matrice augmentée* du système linéaire 3.1 est la matrice de dimension n sur $n+1$ que l'on obtient en ajoutant le membre de droite $\vec{b}$ à la matrice A, c'est-à-dire:

$$\left[\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} & b_2 \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} & b_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} & b_n \end{array}\right] \qquad (3.11)$$

Puisque les opérations élémentaires doivent être effectuées à la fois sur les lignes de la matrice A et sur celles du vecteur $\vec{b}$, cette notation est très utile.

Remarque 3.13
Il arrive également que l'on doive résoudre des systèmes de la forme $A\vec{x} = \vec{b}$ avec k seconds membres $\vec{b}$ différents (la matrice A étant fixée). On peut alors construire la matrice augmentée contenant les k seconds membres désirés. La matrice augmentée ainsi obtenue est de dimension $n \times (n + k)$. □

Exemple 3.8
Considérons l'exemple suivant:

$$\left[\begin{array}{ccc|c} \boxed{2} & 1 & 2 & 10 \\ 6 & 4 & 0 & 26 \\ 8 & 5 & 1 & 35 \end{array}\right] \begin{array}{l} \\ T_1(\vec{l}_2 \leftarrow \vec{l}_2 - (6/\boxed{2})\vec{l}_1) \\ T_2(\vec{l}_3 \leftarrow \vec{l}_3 - (8/\boxed{2})\vec{l}_1) \end{array}$$

On a indiqué ci-dessus la matrice augmentée de même que les opérations élémentaires (et les matrices associées) qui sont nécessaires pour éliminer les termes non nuls sous la diagonale de la première colonne. Il est à noter que l'on divise par 2 (a_{11}) les coefficients qui multiplient la ligne 1. On dit alors que 2 est le *pivot*. On obtient, en effectuant les opérations indiquées:

$$\left[\begin{array}{ccc|c} 2 & 1 & 2 & 10 \\ 0 & \boxed{1} & -6 & -4 \\ 0 & 1 & -7 & -5 \end{array}\right] \begin{array}{l} \\ \\ T_3(\vec{l}_3 \leftarrow \vec{l}_3 - (1/\boxed{1})\vec{l}_2) \end{array}$$

Pour produire une matrice triangulaire supérieure, il suffit maintenant d'introduire des 0 sous la diagonale de la deuxième colonne. L'opération est indiquée ci-dessus et le pivot est 1 puisque maintenant $a_{22} = 1$. On obtient donc:

$$\left[\begin{array}{ccc|c} 2 & 1 & 2 & 10 \\ 0 & 1 & -6 & -4 \\ 0 & 0 & -1 & -1 \end{array}\right] \tag{3.12}$$

Il reste ensuite à faire la remontée triangulaire de l'algorithme 3.4. On obtient:

$$x_3 = -1/-1 = 1$$

d'où:

$$x_2 = \frac{-4 - (-6)(1)}{1} = 2$$

et enfin:

$$x_1 = \frac{10 - (1)(2) - (2)(1)}{2} = 3$$

On a construit le système triangulaire 3.12 en effectuant des opérations élémentaires directement sur les lignes de la matrice. La matrice triangulaire obtenue est notée U. Les opérations effectuées pour obtenir U sont équivalentes à multiplier le système de départ par une suite de matrices inversibles. On a en fait:

$$U = T_3 T_2 T_1 A$$

où les matrices T_i correspondent aux différentes opérations effectuées sur les lignes de la matrice. Plus explicitement, on a:

$$\begin{bmatrix} 2 & 1 & 2 \\ 0 & 1 & -6 \\ 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 2 \\ 6 & 4 & 0 \\ 8 & 5 & 1 \end{bmatrix}$$

Si on poursuit le raisonnement, on a également:

$$A = T_1^{-1} T_2^{-1} T_3^{-1} U$$

Puisque l'on sait inverser les matrices T_i, on a immédiatement que:

$$\begin{bmatrix} 2 & 1 & 2 \\ 6 & 4 & 0 \\ 8 & 5 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 2 \\ 0 & 1 & -6 \\ 0 & 0 & -1 \end{bmatrix}$$

ou encore

$$\begin{bmatrix} 2 & 1 & 2 \\ 6 & 4 & 0 \\ 8 & 5 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 4 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 2 \\ 0 & 1 & -6 \\ 0 & 0 & -1 \end{bmatrix}$$

On remarque que les coefficients de la matrice triangulaire inférieure sont ceux qui ont permis d'éliminer les termes non nuls sous la diagonale de la matrice A. Tout cela revient à décomposer la matrice A en un produit d'une matrice triangulaire inférieure, notée L, et d'une matrice triangulaire supérieure U. C'est ce que l'on appelle une *décomposition LU*.

◆◆◆◆◆

Remarque 3.14
La méthode d'élimination de Gauss revient à factoriser la matrice A en un produit de deux matrices triangulaires L et U *seulement dans le cas où aucune permutation de lignes n'est effectuée.* □

Remarque 3.15
Le déterminant de la matrice de départ est le même que celui de la matrice triangulaire 3.12 puisqu'on n'a effectué que des opérations de la forme $(\vec{l}_i \leftarrow \vec{l}_i + \lambda \vec{l}_j)$, ce qui revient à multiplier le système de départ par une matrice dont le déterminant est 1. On a donc:

$$\text{dét } A = (2)(1)(-1) = -2$$

soit le produit des termes diagonaux de la matrice 3.12. Pour être plus précis:

$$\text{dét } A = \text{dét } T_1^{-1} \text{dét } T_2^{-1} \text{dét } T_3^{-1} \text{dét } U = (1)(1)(1)[(2)(1)(-1)]$$

puisque le déterminant des trois matrices T_i est 1. □

Exemple 3.9
Soit le système linéaire suivant:

$$\begin{array}{rcrcrcrcr} x_1 & - & x_2 & + & 2x_3 & - & x_4 & = & -8 \\ 2x_1 & - & 2x_2 & + & 3x_3 & - & 3x_4 & = & -20 \\ x_1 & + & x_2 & + & x_3 & & & = & -2 \\ x_1 & - & x_2 & + & 4x_3 & + & 3x_4 & = & +4 \end{array}$$

dont la matrice augmentée est:

$$\left[\begin{array}{rrrr|r} \boxed{1} & -1 & 2 & -1 & -8 \\ 2 & -2 & 3 & -3 & -20 \\ 1 & 1 & 1 & 0 & -2 \\ 1 & -1 & 4 & 3 & 4 \end{array}\right] \begin{array}{l} \\ T_1(\vec{l}_2 \leftarrow \vec{l}_2 - (2/\boxed{1})\vec{l}_1) \\ T_2(\vec{l}_3 \leftarrow \vec{l}_3 - (1/\boxed{1})\vec{l}_1) \\ T_3(\vec{l}_4 \leftarrow \vec{l}_4 - (1/\boxed{1})\vec{l}_1) \end{array}$$

En faisant les opérations indiquées (le pivot a_{11} est 1), on élimine les termes non nuls sous la diagonale de la première colonne et on obtient:

$$\left[\begin{array}{rrrr|r} 1 & -1 & 2 & -1 & -8 \\ 0 & \boxed{0} & -1 & -1 & -4 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right] \begin{array}{l} \\ P_4(\vec{l}_2 \leftrightarrow \vec{l}_3) \\ \\ \\ \end{array}$$

Ici, la procédure est interrompue par le fait que le nouveau pivot serait 0 et qu'il n'est pas possible d'éliminer les termes sous ce pivot. Mais on peut encore, parmi les opérations élémentaires, interchanger deux lignes. Le seul choix possible dans cet exemple est d'intervertir la ligne 2 et la ligne 3. On se rend immédiatement compte qu'il n'y a plus que des 0 sous le nouveau pivot et que l'on peut passer à la colonne suivante.

$$\left[\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & \boxed{-1} & -1 & -4 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right] \quad T_5(\vec{l}_4 \leftarrow \vec{l}_4 - (2/\boxed{-1})\vec{l}_3)$$

En effectuant cette dernière opération, on obtient le système triangulaire:

$$\left[\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 0 & 0 & 2 & 4 \end{array}\right]$$

La remontée triangulaire (laissée en exercice) donne la solution:

$$\vec{x} = [-7 \ \ 3 \ \ 2 \ \ 2]^T$$

Encore ici, la matrice triangulaire est le résultat du produit des opérations élémentaires:

$$U = T_5 P_4 T_3 T_2 T_1 A$$

ou encore

$$A = T_1^{-1} T_2^{-1} T_3^{-1} P_4^{-1} T_5^{-1} U$$

qui s'écrit:

$$\left[\begin{array}{cccc} 1 & -1 & 2 & -1 \\ 2 & -2 & 3 & -3 \\ 1 & 1 & 1 & 0 \\ 1 & -1 & 4 & 3 \end{array}\right] = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & -2 & 1 \end{array}\right] \left[\begin{array}{cccc} 1 & -1 & 2 & -1 \\ 0 & 2 & -1 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 2 \end{array}\right]$$

On remarque que la première matrice du terme de droite *n'est pas triangulaire inférieure*. Cela est dû au fait que l'on a permuté deux lignes.

● ● ● ● ●

Remarque 3.16
Le déterminant de la matrice A associée à cet exemple est tout simplement:

$$\begin{aligned} \text{dét } A &= \text{dét } T_1^{-1} \text{dét } T_2^{-1} \text{dét } T_3^{-1} \text{dét } P_4^{-1} \text{dét } T_5^{-1} \text{dét } U \\ &= (1)(1)(1)(-1)(1)[(1)(2)(-1)(2)] \\ &= 4 \end{aligned}$$

Le déterminant est donc le produit de la diagonale de la matrice triangulaire à un signe près puisqu'on a permuté une fois 2 lignes et que dét $P_4 = -1$. □

Nous n'insistons pas davantage sur la méthode d'élimination de Gauss puisque nous avons démontré qu'il s'agit d'un cas particulier de décomposition d'une matrice en un produit d'une matrice triangulaire inférieure et d'une matrice triangulaire supérieure ($A = LU$). Nous abordons maintenant directement cette décomposition.

3.5 Décomposition *LU*

3.5.1 Principe de la méthode

Supposons un instant que nous ayons réussi à exprimer la matrice A en un produit de deux matrices triangulaires L et U. Comment cela nous permet-il de résoudre le système $A\vec{x} = \vec{b}$? Il suffit de remarquer que:

$$A\vec{x} = LU\vec{x} = \vec{b}$$

et de poser $U\vec{x} = \vec{y}$. La résolution du système linéaire se fait alors en deux étapes:

$$\begin{aligned} L\vec{y} &= \vec{b} \\ U\vec{x} &= \vec{y} \end{aligned} \tag{3.13}$$

qui sont deux systèmes triangulaires. On utilise d'abord une descente triangulaire sur la matrice L pour obtenir $\vec{y}$ et par la suite une remontée triangulaire sur la matrice U pour obtenir la solution recherchée $\vec{x}$.

Il faut tout de suite souligner que la décomposition LU n'est pas unique. On peut en effet écrire un nombre réel comme le produit de deux autres nombres d'une infinité de façons. Il en est de même pour les matrices.

Exemple 3.10
Pour illustrer la non-unicité de la décomposition LU, il suffit de vérifier les égalités:

$$\begin{bmatrix} 2 & -1 & -1 \\ 0 & -4 & 2 \\ 6 & -3 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -4 & 0 \\ 6 & 0 & 4 \end{bmatrix} \begin{bmatrix} 1 & -0{,}5 & -0{,}5 \\ 0 & 1 & -0{,}5 \\ 0 & 0 & 1 \end{bmatrix}$$

et:

$$\begin{bmatrix} 2 & -1 & -1 \\ 0 & -4 & 2 \\ 6 & -3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & -1 \\ 0 & -4 & 2 \\ 0 & 0 & 4 \end{bmatrix}$$

● ● ● ● ●

Remarque 3.17
La décomposition LU n'étant pas unique, il faut faire au préalable des choix arbitraires. Le choix le plus populaire consiste à imposer que la matrice U ait des 1 sur sa diagonale. C'est la ***décomposition de Crout***. □

3.5.2 Décomposition de Crout

Pour obtenir cette décomposition (ou *factorisation*), nous considérons une matrice de dimension 4 sur 4, le cas général étant similaire. On doit donc déterminer les coefficients l_{ij} et u_{ij} des matrices L et U de telle sorte que $A = LU$. En imposant que la diagonale de U soit composée de 1, on doit avoir:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} & u_{14} \\ 0 & 1 & u_{23} & u_{24} \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Il suffit de procéder de façon systématique par identification des coefficients. On remarque d'abord qu'il y a exactement 16 (n^2 dans le cas général) inconnues à déterminer. On peut faire le produit des matrices L et U, et se servir des différents coefficients a_{ij}. On obtient ainsi les 16 (n^2) équations nécessaires pour déterminer les coefficients l_{ij} et u_{ij}.

1. *Produit des lignes de L par la première colonne de U*

   On obtient immédiatement que:

$$l_{11} = a_{11} \quad l_{21} = a_{21} \quad l_{31} = a_{31} \quad l_{41} = a_{41}$$

   et la première colonne de L est tout simplement la première colonne de A.

2. *Produit de la première ligne de L par les colonnes de U*

   On obtient respectivement:

$$l_{11}u_{12} = a_{12} \quad l_{11}u_{13} = a_{13} \quad l_{11}u_{14} = a_{14}$$

   d'où on tire que:

$$u_{12} = \frac{a_{12}}{l_{11}} \quad u_{13} = \frac{a_{13}}{l_{11}} \quad u_{14} = \frac{a_{14}}{l_{11}}$$

   On a donc la première ligne de U, si $l_{11} \neq 0$.

3. *Produit des lignes de L par la deuxième colonne de U*

   Les différents produits donnent:

$$\begin{aligned} l_{21}u_{12} + l_{22} &= a_{22} \\ l_{31}u_{12} + l_{32} &= a_{32} \\ l_{41}u_{12} + l_{42} &= a_{42} \end{aligned}$$

   ou encore

$$\begin{aligned} l_{22} &= a_{22} - l_{21}u_{12} \\ l_{32} &= a_{32} - l_{31}u_{12} \\ l_{42} &= a_{42} - l_{41}u_{12} \end{aligned}$$

   et la deuxième colonne de L est connue.

4. *Produit de la deuxième ligne de L par les colonnes de U*

   On trouve immédiatement que:

$$\begin{aligned} l_{21}u_{13} + l_{22}u_{23} &= a_{23} \\ l_{21}u_{14} + l_{22}u_{24} &= a_{24} \end{aligned}$$

ce qui donne:

$$\begin{aligned} u_{23} &= \frac{a_{23} - l_{21}u_{13}}{l_{22}} \\ u_{24} &= \frac{a_{24} - l_{21}u_{14}}{l_{22}} \end{aligned}$$

5. *Produit des lignes de L par la troisième colonne de U*

   La même suite d'opérations donne:

$$\begin{aligned} l_{31}u_{13} + l_{32}u_{23} + l_{33} &= a_{33} \\ l_{41}u_{13} + l_{42}u_{23} + l_{43} &= a_{43} \end{aligned}$$

   ce qui permet d'obtenir la troisième colonne de L:

$$\begin{aligned} l_{33} &= a_{33} - l_{31}u_{13} - l_{32}u_{23} \\ l_{43} &= a_{43} - l_{41}u_{13} - l_{42}u_{23} \end{aligned}$$

6. *Produit de la troisième ligne de L par la quatrième colonne de U*

   On voit que:

$$l_{31}u_{14} + l_{32}u_{24} + l_{33}u_{34} = a_{34}$$

   ce qui permet d'obtenir:

$$u_{34} = \frac{a_{34} - l_{31}u_{14} - l_{32}u_{24}}{l_{33}}$$

7. *Produit de la quatrième ligne de L par la quatrième colonne de U*

   On obtient:

$$l_{41}u_{14} + l_{42}u_{24} + l_{43}u_{34} + l_{44} = a_{44}$$

   Le dernier coefficient recherché est donc:

$$l_{44} = a_{44} - l_{41}u_{14} - l_{42}u_{24} - l_{43}u_{34}$$

De façon générale, on a l'algorithme suivant.

Algorithme 3.1: Décomposition de Crout

1. *Décomposition LU* (sans permutation de lignes)

    - Première colonne de L:

    $$l_{i1} = a_{i1} \quad \text{pour } i = 1, 2, \cdots, n \tag{3.14}$$

    - Première ligne de U:

    $$u_{1i} = \frac{a_{1i}}{l_{11}} \quad \text{pour } i = 2,\ 3, \cdots, n \tag{3.15}$$

    - Pour $i = 2,\ 3,\ 4, \cdots, n-1$:
        - Calcul du pivot:

        $$l_{ii} = a_{ii} - \sum_{k=1}^{i-1} l_{ik} u_{ki} \tag{3.16}$$

        - Pour $j = i+1,\ i+2, \cdots, n$:
            * Calcul de la i^e colonne de L:

            $$l_{ji} = a_{ji} - \sum_{k=1}^{i-1} l_{jk} u_{ki} \tag{3.17}$$

            * Calcul de la i^e ligne de U:

            $$u_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}}{l_{ii}} \tag{3.18}$$

    - Calcul de l_{nn}:

    $$l_{nn} = a_{nn} - \sum_{k=1}^{n-1} l_{nk} u_{kn} \tag{3.19}$$

2. *Descente et remontée triangulaires*

- Descente triangulaire pour résoudre $L\vec{y} = \vec{b}$:
  - $y_1 = b_1/l_{11}$
  - Pour $i = 2, 3, 4, \cdots, n$:

$$y_i = \frac{b_i - \sum_{k=1}^{i-1} l_{ik} y_k}{l_{ii}} \tag{3.20}$$

- Remontée triangulaire pour résoudre $U\vec{x} = \vec{y}$ ($u_{ii} = 1$):
  - $x_n = y_n$
  - Pour $i = n-1, n-2, \cdots 2, 1$:

$$x_i = y_i - \sum_{k=i+1}^{n} u_{ik} x_k \qquad \square \tag{3.21}$$

Remarque 3.18
L'algorithme précédent ne fonctionne que si les pivots l_{ii} sont tous non nuls. Ce n'est pas toujours le cas et il est possible qu'il faille permuter deux lignes pour éviter cette situation, tout comme pour l'élimination de Gauss. Le coefficient l_{ii} est encore appelé ***pivot***. Nous abordons un peu plus loin les techniques de recherche du meilleur pivot. □

Remarque 3.19
Une fois utilisés, les coefficients de la matrice A ne servent plus à rien. Ils peuvent donc être détruits au fur et à mesure que la décomposition progresse. De fait, on peut les remplacer par les valeurs de l_{ij} ou u_{ij} selon le cas. C'est ce que l'on nomme la ***notation compacte***. La notation compacte évite de garder inutilement en mémoire des matrices de grande taille. □

Définition 3.4
La *notation compacte* de la décomposition LU est la matrice de coefficients:

$$\begin{bmatrix} l_{11} & u_{12} & u_{13} & u_{14} \\ l_{21} & l_{22} & u_{23} & u_{24} \\ l_{31} & l_{32} & l_{33} & u_{34} \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \qquad (3.22)$$

dans le cas d'une matrice de dimension 4 sur 4. La matrice initiale A est tout simplement détruite. ***Les coefficients 1 sur la diagonale de la matrice U ne sont pas indiqués explicitement, mais doivent tout de même être pris en compte.*** De façon plus rigoureuse, la notation compacte revient à mettre en mémoire la matrice:

$$L + U - I$$

et à détruire la matrice A.

Exemple 3.11
Soit le système:

$$\begin{bmatrix} 3 & -1 & 2 \\ 1 & 2 & 3 \\ 2 & -2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 12 \\ 11 \\ 2 \end{bmatrix}$$

que l'on doit décomposer en un produit LU. Pour illustrer la notation compacte, ***on remplace au fur et à mesure les coefficients a_{ij} par les coefficients l_{ij} ou u_{ij}; les cases soulignent que l'élément a_{ij} correspondant a été détruit.***

1. Première colonne de L:

   C'est tout simplement la première colonne de A:

$$\begin{bmatrix} \boxed{3} & -1 & 2 \\ \boxed{1} & 2 & 3 \\ \boxed{2} & -2 & -1 \end{bmatrix}$$

2. Première ligne de U:

   Le pivot de la première ligne est 3. On divise donc la première ligne de A par 3:

$$\begin{bmatrix} \boxed{3} & \boxed{-1/3} & \boxed{2/3} \\ \boxed{1} & 2 & 3 \\ \boxed{2} & -2 & -1 \end{bmatrix}$$

3. Deuxième colonne de L:

   De la relation 3.17, on tire:

$$\begin{aligned} l_{22} &= a_{22} - l_{21}u_{12} \\ &= 2 - (1)(-1/3) \\ &= 7/3 \\ l_{32} &= a_{32} - l_{31}u_{12} \\ &= -2 - (2)(-1/3) \\ &= -4/3 \end{aligned}$$

   On a maintenant:

$$\begin{bmatrix} \boxed{3} & \boxed{-1/3} & \boxed{2/3} \\ \boxed{1} & \boxed{7/3} & 3 \\ \boxed{2} & \boxed{-4/3} & -1 \end{bmatrix}$$

4. Deuxième ligne de U:

   De la relation 3.18, on tire:

$$\begin{aligned} u_{23} &= \frac{a_{23} - l_{21}u_{13}}{l_{22}} \\ &= \frac{3 - (1)(2/3)}{7/3} \\ &= 1 \end{aligned}$$

   La matrice compacte devient:

$$\begin{bmatrix} \boxed{3} & \boxed{-1/3} & \boxed{2/3} \\ \boxed{1} & \boxed{7/3} & \boxed{1} \\ \boxed{2} & \boxed{-4/3} & -1 \end{bmatrix}$$

5. Calcul de l_{33}:

   D'après la relation 3.19, on a:

$$\begin{aligned} l_{33} &= a_{33} - l_{31}u_{13} - l_{32}u_{23} \\ &= -1 - (2)(2/3) - (-4/3)(1) \\ &= -1 \end{aligned}$$

   La matrice compacte est donc:

$$\begin{bmatrix} \boxed{3} & \boxed{-1/3} & \boxed{2/3} \\ \boxed{1} & \boxed{7/3} & \boxed{1} \\ \boxed{2} & \boxed{-4/3} & \boxed{-1} \end{bmatrix}$$

   La matrice de départ A (maintenant détruite) vérifie nécessairement:

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 1 & 7/3 & 0 \\ 2 & -4/3 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1/3 & 2/3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

6. Résolution de $L\vec{y} = \vec{b}$:

   La descente triangulaire donne:

$$\begin{aligned} y_1 &= b_1/l_{11} \\ &= 12/3 \\ &= 4 \end{aligned}$$

$$\begin{aligned} y_2 &= \frac{b_2 - l_{21}y_1}{l_{22}} \\ &= \frac{11 - (1)(4)}{7/3} \\ &= 3 \end{aligned}$$

$$\begin{aligned} y_3 &= \frac{b_3 - l_{31}y_1 - l_{32}y_2}{l_{33}} \\ &= \frac{2 - (2)(4) - (-4/3)(3)}{(-1)} \\ &= 2 \end{aligned}$$

7. Résolution de $U\vec{x} = \vec{y}$:

$$\begin{aligned} x_3 &= y_3 \\ &= 2 \end{aligned}$$

$$\begin{aligned} x_2 &= y_2 - u_{23}x_3 \\ &= 3 - (1)(2) \\ &= 1 \end{aligned}$$

$$\begin{aligned} x_1 &= y_1 - u_{12}x_2 - u_{13}x_3 \\ &= 4 - (-1/3)(1) - (2/3)(2) \\ &= 3 \end{aligned}$$

La solution recherchée est donc $\vec{x} = [3 \ \ 1 \ \ 2]^T$.

● ● ● ● ●

3.5.3 Décomposition LU et permutation de lignes

Comme nous l'avons déjà remarqué, l'algorithme de décomposition LU exige que les pivots l_{ii} soient non nuls. Dans le cas contraire, il faut essayer de permuter deux lignes. Contrairement à la méthode d'élimination de Gauss, la décomposition LU n'utilise le terme de droite $\vec{b}$ qu'à la toute fin, au moment de la descente triangulaire $L\vec{y} = \vec{b}$. Si on permute des lignes, on doit en garder la trace de façon à effectuer les mêmes permutations sur $\vec{b}$. À cette fin, on introduit un vecteur $\vec{O}$ dit *de permutation* qui contient tout simplement la numérotation des équations.

Remarque 3.20
Dans une décomposition LU, la permutation de lignes s'effectue toujours après le calcul de chaque colonne de L. On place en position de pivot *le plus grand terme en valeur absolue de cette colonne (sous le pivot actuel)*, pour des raisons de précision que nous verrons plus loin. □

Illustrons cela par un exemple.

Exemple 3.12
Soit:

$$\begin{bmatrix} 0 & 2 & 1 \\ 1 & 0 & 0 \\ 3 & 0 & 1 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Au départ, le vecteur $\vec{O}$ indique que la numérotation des équations n'a pas encore été modifiée.

1. Première colonne de L:

   Puisqu'il s'agit de la première colonne de A, on a

$$\begin{bmatrix} \boxed{0} & 2 & 1 \\ \boxed{1} & 0 & 0 \\ \boxed{3} & 0 & 1 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

   Le vecteur de permutation n'a pas été modifié, mais on a un pivot nul. On effectue alors l'opération $(\vec{l}_1 \leftrightarrow \vec{l}_3)$. On aurait tout aussi bien pu permuter la ligne 1 et la ligne 2, mais on choisit immédiatement le plus grand pivot possible (en valeur absolue). Le vecteur de permutation est alors modifié:

$$\begin{bmatrix} \boxed{3} & 0 & 1 \\ \boxed{1} & 0 & 0 \\ \boxed{0} & 2 & 1 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

2. Première ligne de U:

   Il suffit de diviser cette ligne par le nouveau pivot 3:

$$\begin{bmatrix} \boxed{3} & \boxed{0} & \boxed{1/3} \\ \boxed{1} & 0 & 0 \\ \boxed{0} & 2 & 1 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

3. Deuxième colonne de L:

   De la relation 3.17, on tire:

$$\begin{aligned} l_{22} &= a_{22} - l_{21}u_{12} \\ &= 0 - (1)(0) = 0 \end{aligned}$$

$$\begin{aligned} l_{32} &= a_{32} - l_{31}u_{12} \\ &= 2 - (0)(0) \\ &= 2 \end{aligned}$$

On a maintenant:

$$\begin{bmatrix} \boxed{3} & \boxed{0} & \boxed{1/3} \\ \boxed{1} & \boxed{0} & 0 \\ \boxed{0} & \boxed{2} & 1 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

et encore un pivot nul, qui oblige à intervertir les lignes 2 et 3 et à modifier $\vec{O}$ en conséquence ($\vec{l}_2 \leftrightarrow \vec{l}_3$):

$$\begin{bmatrix} \boxed{3} & \boxed{0} & \boxed{1/3} \\ \boxed{0} & \boxed{2} & 1 \\ \boxed{1} & \boxed{0} & 0 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

4. Calcul de u_{23}:

   La relation 3.18 mène à:

$$\begin{aligned} u_{23} &= \frac{a_{23} - l_{21}u_{13}}{l_{22}} \\ &= \frac{1 - (0)(1/3)}{2} \\ &= 1/2 \end{aligned}$$

   et la matrice compacte devient:

$$\begin{bmatrix} \boxed{3} & \boxed{0} & \boxed{1/3} \\ \boxed{0} & \boxed{2} & \boxed{1/2} \\ \boxed{1} & \boxed{0} & 0 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

5. Calcul de l_{33}:

   On calcule enfin:

$$\begin{aligned} l_{33} &= a_{33} - l_{31}u_{13} - l_{32}u_{23} \\ &= 0 - (1)(1/3) - (0)(1/2) \\ &= -(1/3) \end{aligned}$$

La décomposition LU de la matrice A est donc:

$$\begin{bmatrix} \boxed{3} & \boxed{0} & \boxed{1/3} \\ \boxed{0} & \boxed{2} & \boxed{1/2} \\ \boxed{1} & \boxed{0} & \boxed{-1/3} \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

Il faut toutefois remarquer que le produit LU donne:

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 0 & -1/3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1/3 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

c'est-à-dire *la matrice A permutée suivant le vecteur $\vec{O}$*. On veut maintenant résoudre:

$$A\vec{x} = \begin{bmatrix} 5 \\ -1 \\ -2 \end{bmatrix}$$

Compte tenu du vecteur $\vec{O}$, on résout d'abord:

$$L\vec{y} = \begin{bmatrix} -2 \\ 5 \\ -1 \end{bmatrix}$$

À noter l'ordre des valeurs dans le membre de droite. La descente triangulaire (laissée en exercice) donne $\vec{y} = [-2/3 \quad 5/2 \quad 1]^T$. Il suffit maintenant d'effectuer la remontée triangulaire:

$$U\vec{x} = \begin{bmatrix} -2/3 \\ 5/2 \\ 1 \end{bmatrix}$$

qui nous donne la solution finale $\vec{x} = [-1 \;\; 2 \;\; 1]^T$.

● ● ● ● ●

Remarque 3.21

Le déterminant de la matrice A de l'exemple précédent est donné par:

$$\text{dét } A = (-1)(-1)[(3)(2)(-1/3)] = -2$$

Comme on a permuté deux lignes deux fois, le déterminant a changé de signe deux fois. □

Cela nous amène au théorème suivant.

Théorème 3.1

On peut calculer le déterminant d'une matrice A à l'aide de la méthode de décomposition LU de la façon suivante:

$$\text{dét } A = (-1)^N \prod_{i=1}^{n} l_{ii} \tag{3.23}$$

où N est le nombre de fois où on a interverti deux lignes. □

Nous avons mentionné que la décomposition LU est une méthode directe, c'est-à-dire que l'on peut prévoir le nombre exact d'opérations arithmétiques nécessaires pour résoudre un système d'équations. On a de fait le résultat suivant (voir Burden et Faires, réf. [2]).

Théorème 3.2

Une décomposition LU pour la résolution d'un système linéaire de dimension n sur n requiert exactement:

$$\frac{n^3 - n}{3} \text{ multiplications/divisions}$$

et

$$\frac{2n^3 - 3n^2 + n}{6} \text{ additions/soustractions}$$

à l'étape de décomposition en un produit LU. De plus, les remontée et descente triangulaires nécessitent:

$$n^2 \text{ multiplications/divisions}$$

et

$$n^2 - n \text{ additions/soustractions}$$

pour un total de:

$$\frac{n^3 + 3n^2 - n}{3} \text{ multiplications/divisions}$$

et

$$\frac{2n^3 + 3n^2 - 5n}{6} \text{ additions/soustractions}$$

Du point de vue informatique, une multiplication (ou une division) est une opération plus coûteuse qu'une simple addition (ou soustraction). C'est donc principalement le nombre de multiplications/divisions qui est important. De plus, on note que si n est grand le nombre total de multiplications/divisions est de l'ordre de $n^3/3$, en négligeant les puissances de n inférieures à 3.

Enfin, et cela est très important, *la décomposition de la matrice A en un produit LU coûte beaucoup plus cher ($\simeq n^3/3$ multiplications/divisions) que les remontée et descente triangulaires ($\simeq n^2$ multiplications/divisions). Le gros du travail se trouve donc dans la décomposition elle-même.* □

3.5.4 Calcul de la matrice inverse A^{-1}

Le calcul de la matrice inverse A^{-1} est rarement nécessaire. En effet, il est inutile de calculer A^{-1} pour résoudre un système linéaire. Nous avons vu dans les sections précédentes comment parvenir à une solution sans jamais faire intervenir A^{-1}. Cependant, si pour une raison ou une autre on souhaite calculer cet inverse, il est important de suivre le bon cheminement afin d'éviter des calculs longs et parfois inutiles.

Nous avons indiqué que la solution du système linéaire 3.2 est donnée par:

$$\vec{x} = A^{-1}\vec{b}$$

Si on veut déterminer la matrice inverse, il suffit de remarquer que le produit d'une matrice par le vecteur $\vec{e}_i$ dont toutes les composantes sont nulles, sauf la i^e qui vaut 1, donne la i^e colonne de la matrice A. L'exemple suivant en fait la démonstration.

Exemple 3.13

La produit de la matrice A suivante par le vecteur $\vec{e}_3$ donne:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix}$$

qui est bien la troisième colonne de la matrice A de départ.

● ● ● ● ●

Si on applique ce raisonnement à la matrice A^{-1}, on constate qu'après avoir noté $\vec{c}_i$, la i^e colonne de A^{-1}, on a:

$$\vec{c}_i = A^{-1}\vec{e}_i$$

ou de façon équivalente

$$A\vec{c}_i = \vec{e}_i \tag{3.24}$$

La résolution de la relation 3.24 donne la i^e colonne de A^{-1}. On peut donc affirmer que le calcul de la matrice A^{-1} est équivalent à la résolution de n systèmes linéaires (un par colonne de A^{-1}).

Remarque 3.22
Puisque le calcul de A^{-1} est équivalent à la résolution de n systèmes linéaires, il est clair qu'il ne faut jamais calculer A^{-1} pour résoudre un système linéaire. Il vaut mieux utiliser directement une décomposition LU sans passer par l'inverse. □

Remarque 3.23
Si on veut quand même calculer A^{-1}, il faut effectuer d'abord la décomposition LU de A *une seule fois* ($\simeq n^3/3$ multiplications/divisions), puis n remontées et descentes triangulaires ($\simeq n \times n^2 = n^3$ multiplications/divisions), pour un total approximatif de $4n^3/3$ multiplications/divisions. Ces évaluations montrent bien que le calcul d'un inverse est beaucoup plus coûteux ($\simeq 4n^3/3$ multiplications/divisions) que la résolution d'un système linéaire ($\simeq n^3/3$ multiplications/divisions).

Le tableau suivant indique le nombre approximatif de multiplications et de divisions nécessaires pour résoudre un système linéaire $A\vec{x} = \vec{b}$ et pour calculer l'inverse de la matrice A, pour différentes valeurs de n.

n	Résolution de $A\vec{x} = \vec{b}$ ($\simeq n^3/3$ mult./div.)	Calcul deA^{-1} ($\simeq 4n^3/3$ mult./div.)
10	333	1 333
100	333 333	1 333 333
1000	333 333 333	1 333 333 333

Ainsi, le calcul de l'inverse d'une matrice de dimension 1000 sur 1000 nécessite à peu près un milliard de multiplications/divisions de plus que la résolution d'un système linéaire de même dimension. □

Exemple 3.14
On doit calculer l'inverse de la matrice:

$$\begin{bmatrix} 0 & 2 & 1 \\ 1 & 0 & 0 \\ 3 & 0 & 1 \end{bmatrix}$$

dont nous avons déjà obtenu la décomposition LU:

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 0 & -1/3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1/3 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

On a recours encore une fois au vecteur de permutation $\vec{O}$. Pour obtenir la matrice inverse de A, on doit résoudre les trois systèmes linéaires suivants:

$$A\vec{c}_1 = \vec{e}_1 \qquad A\vec{c}_2 = \vec{e}_2 \qquad A\vec{c}_3 = \vec{e}_3$$

dont le résultat nous donne les trois colonnes de la matrice A^{-1}. Le premier système est résolu d'abord par la descente triangulaire:

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 0 & -1/3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Il faut prendre garde ici au membre de droite. Il s'agit bien du vecteur $\vec{e}_1 = [1 \quad 0 \quad 0]^T$, mais ordonné suivant le vecteur $\vec{O} = [3 \quad 1 \quad 2]^T$ pour tenir compte des lignes qui ont été permutées lors de la décomposition LU. La résolution conduit à $\vec{y} = [0 \quad 1/2 \quad 0]^T$. Il reste à effectuer la remontée triangulaire:

$$\begin{bmatrix} 1 & 0 & 1/3 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/2 \\ 0 \end{bmatrix}$$

dont le résultat $[0 \quad 1/2 \quad 0]^T$ représente la première colonne de A^{-1}. Le deuxième système exige dans un premier temps la résolution de:

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 0 & -1/3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

(à surveiller l'ordre des composantes du vecteur $\vec{e}_2$ à droite), dont la solution est $\vec{y} = [0 \;\; 0 \;\; -3]^T$. Par la suite:

$$\begin{bmatrix} 1 & 0 & 1/3 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -3 \end{bmatrix}$$

qui donne la deuxième colonne de A^{-1}, soit $\vec{c}_2 = [1 \;\; 3/2 \;\; -3]^T$. Enfin, un raisonnement similaire déterminerait la troisième colonne $\vec{c}_3 = [0 \;\; -1/2 \;\; 1]^T$. La matrice inverse est donc:

$$A^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 3/2 & -1/2 \\ 0 & -3 & 1 \end{bmatrix}$$

● ● ● ● ●

3.6 Effets de l'arithmétique flottante

Jusqu'ici, nous n'avons utilisé que l'arithmétique exacte. Il est grandement temps de regarder si l'arithmétique flottante utilisée par les ordinateurs a une influence quelconque sur les résultats. Il est probable que oui. En fait, nous allons voir que certaines matrices sont très sensibles aux effets de l'arithmétique flottante et d'autres, très peu. Dans le cas de matrices sensibles, nous parlerons de *matrices mal conditionnées.*

Remarque 3.24
En arithmétique flottante à m chiffres dans la mantisse, on doit effectuer chaque opération arithmétique en représentant les opérandes en notation flottante et en arrondissant le résultat de l'opération au m^e *chiffre de la mantisse* (voir la section 1.5). □

Illustrons cela à l'aide d'un exemple.

Exemple 3.15

On doit effectuer la décomposition LU en arithmétique flottante à 4 chiffres de la matrice suivante:

$$\begin{bmatrix} 1{,}012 & -2{,}132 & 3{,}104 \\ -2{,}132 & 4{,}096 & -7{,}013 \\ 3{,}104 & -7{,}013 & 0{,}014 \end{bmatrix}$$

Les opérations sont résumées ci-dessous:

$$\begin{aligned} l_{11} &= 1{,}012 \\ l_{21} &= -2{,}132 \\ l_{31} &= 3{,}104 \end{aligned}$$

$$\begin{aligned} u_{12} &= \text{fl}(-2{,}132/1{,}012) \\ &= -2{,}107 \end{aligned}$$

$$\begin{aligned} u_{13} &= \text{fl}(3{,}104/1{,}012) \\ &= 3{,}067 \end{aligned}$$

$$\begin{aligned} l_{22} &= \text{fl}(4{,}096 - \text{fl}[(-2{,}132)(-2{,}107)]) \\ &= \text{fl}(4{,}096 - 4{,}492) \\ &= -0{,}3960 \end{aligned}$$

$$\begin{aligned} l_{32} &= \text{fl}(-7{,}013 - \text{fl}[(3{,}104)(-2{,}107)]) \\ &= \text{fl}(-7{,}013 + 6{,}540) \\ &= -0{,}4730 \end{aligned}$$

$$\begin{aligned} u_{23} &= \text{fl}\left(\frac{-7{,}013 - \text{fl}[(-2{,}132)(3{,}067)]}{-0{,}3960}\right) \\ &= \text{fl}\left(\frac{-7{,}013 + 6{,}539}{-0{,}3960}\right) \\ &= 1{,}197 \end{aligned}$$

$$\begin{aligned} l_{33} &= \text{fl}(0{,}0140 - \text{fl}[(3{,}104)(3{,}067)] - \text{fl}[(-0{,}4730)(1{,}197)]) \\ &= \text{fl}(0{,}0140 - 9{,}520 + 0{,}5662) \\ &= -8{,}940 \end{aligned}$$

On a donc:

$$LU = \begin{bmatrix} 1{,}012 & 0 & 0 \\ -2{,}132 & -0{,}3960 & 0 \\ 3{,}104 & -0{,}4730 & -8{,}940 \end{bmatrix} \begin{bmatrix} 1 & -2{,}107 & 3{,}067 \\ 0 & 1 & 1{,}197 \\ 0 & 0 & 1 \end{bmatrix}$$

●●●●●

Les exemples suivants montrent comment l'arithmétique flottante peut affecter sensiblement la précision de la résolution d'un système linéaire. Nous discutons également des moyens d'en diminuer les effets.

Exemple 3.16
Soit le système suivant:

$$\begin{bmatrix} 1 & 2 \\ 1{,}1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 10 \\ 10{,}4 \end{bmatrix}$$

dont la solution exacte est $\vec{x} = [4 \quad 3]^T$. Si on remplace le terme 1,1 de la matrice par 1,05, la nouvelle solution exacte devient $\vec{x} = [8 \quad 1]^T$. Cet exemple démontre qu'*une petite modification sur un terme de la matrice peut entraîner une grande modification de la solution exacte.* En pratique, l'arithmétique flottante provoque inévitablement de petites modifications de chaque terme de la matrice et de sa décomposition LU. Il est alors tout à fait possible que ces petites erreurs aient d'importantes répercussions sur la solution et, donc, que les résultats numériques soient très éloignés de la solution exacte.

●●●●●

Exemple 3.17
Considérons le système:

$$\begin{bmatrix} 0{,}0003 & 3{,}0000 \\ 1{,}0000 & 1{,}0000 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2{,}0001 \\ 1{,}0000 \end{bmatrix}$$

dont la solution exacte est $\vec{x} = [1/3 \quad 2/3]^T$. Il s'agit maintenant d'effectuer la décomposition LU en arithmétique flottante à 4 chiffres. On remarque que

le système devient en notation flottante à 4 chiffres:

$$\begin{bmatrix} 0{,}3000 \times 10^{-3} & 0{,}3000 \times 10^{1} \\ 0{,}1000 \times 10^{1} & 0{,}1000 \times 10^{1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0{,}2000 \times 10^{1} \\ 0{,}1000 \times 10^{1} \end{bmatrix}$$

et que le 1 de 2,0001 disparaît. La décomposition LU donne dans ce cas:

$$LU = \begin{bmatrix} 0{,}3000 \times 10^{-3} & 0 \\ 0{,}1000 \times 10^{1} & -0{,}9999 \times 10^{4} \end{bmatrix} \begin{bmatrix} 0{,}1000 \times 10^{1} & 0{,}1000 \times 10^{5} \\ 0 & 0{,}1000 \times 10^{1} \end{bmatrix}$$

Le terme u_{12} est très grand puisque le pivot 0,0003 est presque nul. La descente triangulaire donne alors:

$$y_1 = \mathrm{fl}\left(\frac{0{,}2000 \times 10^{1}}{0{,}3000 \times 10^{-3}}\right) = 0{,}6667 \times 10^{4}$$

et

$$y_2 = \mathrm{fl}\left(\frac{1 - 6667}{-9999}\right) = 0{,}6667$$

Puis la remontée triangulaire donne:

$$x_2 = 0{,}6667$$

et

$$x_1 = 6667 - (10\,000)(0{,}6667) = 0$$

Si on compare ce résultat avec la solution exacte $[1/3 \;\; 2/3]^T$, on constate une variation importante de la valeur de x_1. On imagine aisément ce qui peut se produire avec un système de plus grande taille. Mais comment peut-on limiter les dégats? Une première possibilité consiste à utiliser plus de chiffres dans la mantisse, mais cela n'est pas toujours possible. Si on passe en revue les calculs précédents, on en vient rapidement à soupçonner que la source des ennuis est la division par un pivot presque nul. On sait qu'une telle opération est dangereuse numériquement. Une solution de rechange consiste donc à permuter les lignes même si le pivot n'est pas parfaitement nul. Dans notre exemple, on aura:

$$\begin{bmatrix} 1{,}0000 & 1{,}0000 \\ 0{,}0003 & 3{,}0000 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1{,}0000 \\ 2{,}0001 \end{bmatrix}$$

Cette fois, la décomposition LU (toujours à 4 chiffres) donne:

$$LU = \begin{bmatrix} 1{,}0000 & 0 \\ 0{,}0003 & 3{,}0000 \end{bmatrix} \begin{bmatrix} 1{,}0000 & 1{,}0000 \\ 0 & 1{,}0000 \end{bmatrix}$$

car:

$$l_{22} = \text{fl}[3 - \text{fl}[(1)(0{,}0003)]] = \text{fl}[3 - 0{,}0003] = 3$$

La descente triangulaire donne $\vec{y} = [1 \;\; 0{,}6666]^T$ et la remontée nous donne la solution $\vec{x} = [0{,}3334 \;\; 0{,}6666]^T$, qui est très près de la solution exacte.

● ● ● ● ●

Remarque 3.25
Une excellente stratégie de recherche du pivot consiste, une fois la i^e colonne de L calculée, à placer en position de pivot le plus grand terme en valeur absolue de cette colonne. Cette recherche ne tient compte que des lignes situées sous le pivot actuel. □

Exemple 3.18
Cet exemple illustre comment effectuer une permutation de façon systématique. Seules les grandes étapes de la décomposition sont indiquées, les calculs étant laissés en exercice. Les coefficients de la matrice sont détruits au fur et à mesure que les calculs progressent et sont remplacés par l_{ij} ou u_{ij}. Considérons donc la matrice:

$$\begin{bmatrix} 1 & 6 & 9 \\ 2 & 1 & 2 \\ 3 & 6 & 9 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

La première colonne de L étant la première colonne de A, on a

$$\begin{bmatrix} \boxed{1} & 6 & 9 \\ \boxed{2} & 1 & 2 \\ \boxed{3} & 6 & 9 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

On peut alors permuter la ligne 3 et la ligne 1 de manière à placer en position de pivot le plus grand terme de la première colonne de L.

On a maintenant:

$$\begin{bmatrix} \boxed{3} & 6 & 9 \\ \boxed{2} & 1 & 2 \\ \boxed{1} & 6 & 9 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

On calcule la première ligne de U:

$$\begin{bmatrix} \boxed{3} & \boxed{2} & \boxed{3} \\ \boxed{2} & 1 & 2 \\ \boxed{1} & 6 & 9 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

La deuxième colonne de L devient alors:

$$\begin{bmatrix} \boxed{3} & \boxed{2} & \boxed{3} \\ \boxed{2} & \boxed{-3} & 2 \\ \boxed{1} & \boxed{4} & 9 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

On voit qu'il faut maintenant permuter les deux dernières lignes pour amener en position de pivot le plus grand terme de la colonne, qui est 4.

$$\begin{bmatrix} \boxed{3} & \boxed{2} & \boxed{3} \\ \boxed{1} & \boxed{4} & 9 \\ \boxed{2} & \boxed{-3} & 2 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

En continuant ainsi, on trouve la décomposition LU sous forme compacte:

$$\begin{bmatrix} \boxed{3} & \boxed{2} & \boxed{3} \\ \boxed{1} & \boxed{4} & \boxed{3/2} \\ \boxed{2} & \boxed{-3} & \boxed{9/2} \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

● ● ● ● ●

La stratégie de recherche du pivot améliore souvent la précision des résultats, mais cette opération n'est pas toujours suffisante. L'exemple qui suit montre comment la *mise à l'échelle* peut également contribuer à la qualité des résultats.

Exemple 3.19
Soit le système suivant:

$$\begin{bmatrix} 2{,}0000 & 100\,000 \\ 1{,}0000 & 1{,}0000 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 100\,000 \\ 2{,}0000 \end{bmatrix}$$

dont la solution exacte est $[1{,}00002 \;\; 0{,}99998]^T$. Nul besoin ici de rechercher un plus grand pivot. La décomposition LU (en arithmétique flottante à 4 chiffres) donne:

$$LU = \begin{bmatrix} 2 & 0 \\ 1 & -50\,000 \end{bmatrix} \begin{bmatrix} 1 & 50\,000 \\ 0 & 1 \end{bmatrix}$$

La descente triangulaire conduit à $\vec{y} = [50\,000 \;\; 1]^T$ et la remontée triangulaire, à $\vec{x} = [0 \;\; 1]^T$. Ici encore, l'erreur est considérable par rapport à la solution exacte. Cet écart est dû au fait que la matrice A est constituée de termes d'ordre de grandeur très différents. Par exemple, quand on calcule le terme l_{22}, on doit effectuer en arithmétique flottante à 4 chiffres:

$$1 - (1)(50\,000) = -50\,000$$

On a donc effectué une autre opération dangereuse, à savoir soustraire (ou additionner) des termes dont les ordres de grandeur sont très différents. Une solution partielle à ce problème est d'effectuer une mise à l'échelle des coefficients de la matrice.

● ● ● ● ●

Définition 3.5
La *mise à l'échelle* consiste à diviser chaque ligne du système linéaire par le plus grand terme (en valeur absolue) de la ligne correspondante de la matrice A. *On ne tient pas compte du terme de droite $\vec{b}$ pour déterminer le plus grand terme de chaque ligne.*

Dans notre exemple, il suffit de diviser la première ligne par 100 000 (le plus grand terme de la deuxième ligne étant 1) et de résoudre:

$$\begin{bmatrix} 0{,}2000 \times 10^{-4} & 1 \\ 1{,}0000 & 1{,}0000 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2{,}0000 \end{bmatrix}$$

La recherche d'un nouveau pivot est maintenant nécessaire. On peut montrer que la résolution en arithmétique flottante à 4 chiffres donne la solution $\vec{x} = [1 \ \ 1]^T$, ce qui est beaucoup plns près du résultat exact.

Remarque 3.26

La mise à l'échelle utilise la dernière opération élémentaire sur les lignes d'un système linéaire, soit la multiplication d'une ligne par un scalaire. □

Remarque 3.27

Lorsqu'on multiplie une ligne par un scalaire, le déterminant de la matrice est multiplié par ce scalaire. Après avoir effectué la décomposition LU, on peut récupérer le déterminant de la matrice A en divisant le produit de la diagonale de L par ce scalaire. □

Les exemples précédents montrent clairement que certains systèmes linéaires sont très sensibles aux erreurs dues à l'arithmétique flottante. Dans la prochaine section, nous allons essayer de mesurer cette sensibilité.

3.7 Conditionnement d'une matrice

Cette section traite d'erreur et de mesure d'erreur liée aux systèmes linéaires. Il nous faut tout de suite introduire une métrique permettant de mesurer l'écart entre une solution numérique et une solution exacte. Cela nous amène donc à aborder la notion de norme vectorielle au sens de la définition suivante.

Définition 3.6
Une *norme vectorielle* est une application de R^n dans R (R désigne l'ensemble des réels) qui associe à un vecteur $\vec{x}$ un scalaire noté $||\vec{x}||$ et qui vérifie les trois propriétés suivantes:

1. La norme d'un vecteur est toujours strictement positive, sauf si le vecteur a toutes ses composantes nulles:

$$||\vec{x}|| > 0, \text{ sauf si } \vec{x} = \vec{0} \tag{3.25}$$

2. Si α est un scalaire, alors:

$$|| \alpha\vec{x} || = |\alpha| \, ||\vec{x}|| \tag{3.26}$$

   où $|\alpha|$ est la valeur absolue de α.

3. L'*inégalité triangulaire* est toujours vérifiée entre deux vecteurs $\vec{x}$ et $\vec{y}$ quelconques:

$$||\vec{x} + \vec{y}|| \leq ||\vec{x}|| + ||\vec{y}|| \tag{3.27}$$

Toute application vérifiant ces trois propriétés est une norme vectorielle. La plus connue est sans doute la *norme euclidienne*.

Définition 3.7
La norme euclidienne d'un vecteur $\vec{x}$ est notée $||\vec{x}||_e$ et est définie par:

$$||\vec{x}||_e = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{3.28}$$

Théorème 3.3
La norme euclidienne vérifie les trois propriétés d'une norme vectorielle.

Démonstration (facultative):

1. La démonstration de la propriété 3.25 est triviale puisque la racine carrée d'un nombre ne peut s'annuler que si ce nombre est nul. Or,

pour que le terme sous la racine carrée soit nul, il faut que toutes les composantes de $\vec{x}$ soient nulles.

2. La propriété 3.26 découle de:

$$\begin{aligned} ||\alpha\vec{x}||_e &= \sqrt{\alpha^2 x_1^2 + \alpha^2 x_2^2 + \alpha^2 x_3^2 + \cdots + \alpha^2 x_n^2} \\ &= |\alpha|\sqrt{x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2} \\ &= |\alpha|\, ||\vec{x}||_e \end{aligned}$$

3. L'inégalité triangulaire 3.27 est un peu plus difficile à obtenir. Soit $\vec{x}$ et $\vec{y}$, deux vecteurs et α, un scalaire. En vertu de la propriété 3.25 déjà démontrée, on a d'abord pour un scalaire α quelconque:

$$\begin{aligned} 0 &\leq ||\vec{x} + \alpha\vec{y}||_e^2 \\ &= ((x_1 + \alpha y_1)^2 + (x_2 + \alpha y_2)^2 + \cdots + (x_n + \alpha y_n)^2) \\ &= (x_1^2 + x_2^2 + \cdots + x_n^2) + 2\alpha(x_1y_1 + x_2y_2 + \cdots + x_ny_n) \\ &\quad + \alpha^2(y_1^2 + y_2^2 + \cdots + y_n^2) \\ &= ||\vec{x}||_e^2 + 2\alpha(\vec{x}\cdot\vec{y}) + \alpha^2||\vec{y}||_e^2 \\ &= C + B\alpha + A\alpha^2 \end{aligned}$$

c'est-à-dire un polynôme du deuxième degré en α. On aura reconnu dans l'expression précédente le *produit scalaire* habituel de deux vecteurs, noté $(\vec{x}\cdot\vec{y})$. Puisque ce polynôme est toujours positif, et ce quel que soit α, le discriminant $B^2 - 4AC$ doit vérifier l'inégalité:

$$B^2 - 4AC \leq 0 \qquad (3.29)$$

Dans le cas contraire, le polynôme aurait deux racines réelles et prendrait des valeurs négatives entre ces racines. En remplaçant les valeurs de A, B et C dans la relation 3.29 et en divisant par 4, on obtient:

$$|\vec{x}\cdot\vec{y}| \leq ||\vec{x}||_e\, ||\vec{y}||_e \qquad (3.30)$$

qui est l'*inégalité de Cauchy*. Cette inégalité permet de démontrer la troisième propriété. En effet, en prenant $\alpha = 1$ et en utilisant l'inégalité de Cauchy, on a:

$$\begin{aligned} ||\vec{x} + \vec{y}||_e^2 &= ||\vec{x}||_e^2 + 2(\vec{x} \cdot \vec{y}) + ||\vec{y}||_e^2 \\ &\leq ||\vec{x}||_e^2 + 2||\vec{x}||_e^2\, ||\vec{y}||_e + ||\vec{y}||_e^2 \\ &= (||\vec{x}||_e + ||\vec{y}||_e)^2 \end{aligned}$$

qui est le résultat attendu. □

On peut définir, en plus de la norme euclidienne, plusieurs normes vérifiant les trois propriétés nécessaires.

Définition 3.8 *Normes l_1 et l_∞*
La norme l_1 est définie par:

$$||\vec{x}||_1 = \sum_{i=1}^{n} |x_i| \qquad (3.31)$$

tandis que la norme l_∞ est définie par:

$$||\vec{x}||_\infty = \max_{1 \leq i \leq n} |x_i| \qquad (3.32)$$

Exemple 3.20
Pour le vecteur $\vec{x} = [1 \quad -3 \quad -8]^T$, on a:

$$\begin{aligned} ||\vec{x}||_1 &= 1 + 3 + 8 &&= 12 \\ ||\vec{x}||_\infty &= \max(1, 3, 8) &&= 8 \\ ||\vec{x}||_e &= \sqrt{1 + 9 + 64} &&= \sqrt{74} \end{aligned}$$

● ● ● ● ●

Puisque nous nous intéressons plus particulièrement aux systèmes linéaires, il importe de pouvoir définir des normes relatives aux matrices.

Définition 3.9
Une *norme matricielle* est une application qui associe à une matrice A un scalaire noté $||A||$ vérifiant les quatre propriétés suivantes:

1. La norme d'une matrice est toujours strictement positive, sauf si la matrice a toutes ses composantes nulles:

$$||A|| > 0, \text{ sauf si } A = 0 \tag{3.33}$$

2. Si α est un scalaire, alors:

$$|| \alpha A ||= |\alpha| \; ||A|| \tag{3.34}$$

3. L'*inégalité triangulaire* est toujours vérifiée entre deux matrices A et B quelconques, c'est-à-dire:

$$||A + B|| \leq ||A|| + ||B|| \tag{3.35}$$

4. Une quatrième propriété est nécessaire pour les matrices:

$$||AB|| \leq ||A|| \; ||B|| \tag{3.36}$$

Toute application qui vérifie ces quatre propriétés est une norme matricielle. Voici quelques exemples.

Définition 3.10 *Quelques normes matricielles*

$$||A||_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|$$

$$||A||_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$$

$$||A||_2 = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}$$

La norme $||A||_1$ consiste à sommer (en valeur absolue) chacune des colonnes de A et à choisir la plus grande somme. La norme $||A||_\infty$ fait un travail similaire sur les lignes. Enfin, la norme $||A||_2$ est en quelque sorte l'équivalent de la norme euclidienne pour les matrices. On l'appelle quelquefois la *norme de Frobenius.*

Exemple 3.21
Soit la matrice:

$$\begin{bmatrix} 1 & -2 & 5 \\ -3 & 1 & -5 \\ 1 & -9 & 0 \end{bmatrix}$$

Les différentes normes prennent alors les valeurs suivantes:

$$||A||_1 = \max(5, 12, 10) = 12$$

$$||A||_\infty = \max(8, 9, 10) = 10$$

$$||A||_2 = \sqrt{1 + 4 + 25 + 9 + 1 + 25 + 1 + 81} = \sqrt{147}$$

● ● ● ● ●

Il ne reste qu'un point important à aborder pour avoir un portrait complet de la situation. Nous avons des normes vectorielles et matricielles qui permettent de manipuler des vecteurs et des matrices respectivement. Lorsque

l'on s'intéresse aux systèmes linéaires, on doit souvent manipuler des produits de matrices par des vecteurs, d'où l'intérêt de la définition suivante.

Définition 3.11
Une norme vectorielle et une norme matricielle sont dites *compatibles* si la condition:

$$||A\vec{x}|| \leq ||A|| \, ||\vec{x}|| \tag{3.37}$$

est valide quels que soient la matrice A et le vecteur $\vec{x}$.

Les trois normes qui apparaissent dans la relation 3.37 sont respectivement une norme vectorielle (car $A\vec{x}$ est un vecteur), une norme matricielle et une norme vectorielle.

Remarque 3.28
Les normes vectorielles et matricielles ne sont pas toutes compatibles entre elles. On peut démontrer que:

$$||\vec{x}||_1 \quad \text{et} \quad ||A||_1$$

$$||\vec{x}||_\infty \quad \text{et} \quad ||A||_\infty$$

$$||\vec{x}||_e \quad \text{et} \quad ||A||_2$$

sont compatibles deux à deux. □

Exemple 3.22
Considérons de nouveau le vecteur $\vec{x} = [1 \quad -3 \quad -8]^T$ et la matrice:

$$\begin{bmatrix} 1 & -2 & 5 \\ -3 & 1 & -5 \\ 1 & -9 & 0 \end{bmatrix}$$

Le produit $A\vec{x}$ donne le vecteur $[-33 \ \ 34 \ \ 28]^T$ et donc:

$$||A\vec{x}||_1 = 95 \quad ||A\vec{x}||_\infty = 34 \quad \text{et} \quad ||A\vec{x}||_e = \sqrt{3029}$$

L'inégalité 3.37 devient respectivement en norme l_1:

$$95 \leq (12)(12)$$

en norme l_∞:

$$34 \leq (10)(8)$$

et en norme euclidienne:

$$\sqrt{3029} \leq (\sqrt{147})(\sqrt{74}) = \sqrt{10\,878}$$

●●●●●

Nous en arrivons au point clé de cette section, qui est le conditionnement d'nne matrice. Introduisons d'abord sa définition.

Définition 3.12
Le conditionnement d'une matrice (noté cond$A$) est défini par:

$$\mathrm{cond}A = ||A||\,||A^{-1}|| \tag{3.38}$$

Il s'agit simplement du produit de la norme de A et de la norme de son inverse.

Remarque 3.29
Le conditionnement dépend de la norme matricielle utilisée. On utilise le plus souvent la norme $||A||_\infty$. □

Il ne reste plus qu'à montrer en quoi le conditionnement d'une matrice est si important pour déterminer la sensibilité d'une matrice aux erreurs d'arrondis et à l'arithmétique flottante.

Tout d'abord, on montre que le conditiounement est un nombre supérieur ou égal à 1. En effet, si I désigne la matrice identité (ayant des 1 sur la diagonale et des zéros partout ailleurs), on a:

$$||A|| = ||AI|| \leq ||A||\,||I||$$

en vertu de la relation 3.36. Cela entraîne, après division par $||A||$ de chaque côté, que $||I|| \geq 1$, quelle que soit la norme matricielle utilisée. On en conclut que:

$$1 \leq ||I|| = ||AA^{-1}|| \leq ||A||\,||A^{-1}||$$

et donc que:

$$1 \leq \text{cond}A < \infty \tag{3.39}$$

3.7.1 Bornes d'erreurs et conditionnement

Considérons le système linéaire:

$$A\vec{x} = \vec{b}$$

et notons $\vec{x}$, la solution exacte et $\vec{x}^*$, une solution approximative qu'on obtient en utilisant l'arithmétique flottante. Ces deux vecteurs devraient être près l'un de l'autre, c'est-à-dire que la norme de l'erreur:

$$||\vec{e}|| = ||\vec{x} - \vec{x}^*||$$

devrait être petite. *Ce n'est pas toujours le cas.* Définissons le résidu par:

$$\vec{r} = \vec{b} - A\vec{x}^* \tag{3.40}$$

On a alors:

$$\vec{r} = \vec{b} - A\vec{x}^* = A\vec{x} - A\vec{x}^* = A(\vec{x} - \vec{x}^*) = A\vec{e}$$

ce qui signifie que $\vec{e} = A^{-1}\vec{r}$. Si on utilise des normes vectorielles et matricielles compatibles, on a en vertu de la relation 3.37:

$$||\vec{e}|| \leq ||A^{-1}||\ ||\vec{r}|| \tag{3.41}$$

De façon analogue, puisque $A\vec{e} = \vec{r}$:

$$||\vec{r}|| \leq ||A||\ ||\vec{e}||$$

qui peut s'écrire:

$$\frac{||\vec{r}||}{||A||} \leq ||\vec{e}|| \tag{3.42}$$

En regroupant les relations 3.41 et 3.42, on obtient:

$$\frac{||\vec{r}||}{||A||} \leq ||\vec{e}|| \leq ||A^{-1}||\ ||\vec{r}|| \tag{3.43}$$

Par ailleurs, en refaisant le même raisonnement avec les égalités $A\vec{x} = \vec{b}$ et $\vec{x} = A^{-1}\vec{b}$, on trouve:

$$\frac{||\vec{b}||}{||A||} \leq ||\vec{x}|| \leq ||A^{-1}||\ ||\vec{b}||$$

Après avoir inversé ces inégalités, on trouve:

$$\frac{1}{||A^{-1}||\,||\vec{b}||} \leq \frac{1}{||\vec{x}||} \leq \frac{||A||}{||\vec{b}||} \tag{3.44}$$

En multipliant les inégalités 3.43 et 3.44, on obtient le résultat fondamental suivant.

Théorème 3.4

$$\frac{1}{\mathrm{cond}A}\frac{||\vec{r}||}{||\vec{b}||} \leq \frac{||\vec{e}||}{||\vec{x}||} \leq \mathrm{cond}A\,\frac{||\vec{r}||}{||\vec{b}||} \qquad \square \tag{3.45}$$

Remarque 3.30
Plusieurs remarques s'imposent pour bien comprendre l'inégalité précédente.

1. Le terme du milieu représente l'erreur relative entre la solution exacte $\vec{x}$ et la solution approximative $\vec{x}^*$.

2. Si le conditionnement de la matrice A est près de 1, l'erreur relative est coincée entre deux valeurs très près l'une de l'autre. Si la norme du résidu est petite, l'erreur relative est également petite et la précision de la solution approximative a toutes le chances d'être satisfaisante.

3. Par contre, si le conditionnement de la matrice A est grand, la valeur de l'erreur relative est quelque part entre 0 et un nombre possiblement très grand. *Il est donc à craindre que l'erreur relative soit alors grande, donc que la solution approximative soit de faible précision et même, dans certains cas, complètement fausse.*

4. *Même si la norme du résidu est petite, il est possible que l'erreur relative liée à la solution approximative soit quand même très grande.*

5. Plus le conditionnement de la matrice A est grand, plus on doit être attentif à l'algorithme de résolution utilisé.

6. Il importe de rappeler que, même si une matrice est bien conditionnée, un mauvais algorithme de résolution peut conduire à des résultats erronés. $\square$

On peut obtenir une autre inégalité qui illustre le rôle du conditionnement d'une matrice quant à la précision de la solution numérique d'un système linéaire. Soit le système linéaire:

$$A\vec{x} = \vec{b}$$

Lorsque l'on résout un tel système sur ordinateur, où la représentation des nombres n'est pas toujours exacte, on résout en fait:

$$(A + E)\vec{x}^* = \vec{b}$$

où la matrice E représente une perturbation du système initial, due par exemple aux erreurs de représentation sur ordinateur des coefficients de la matrice A. La matrice E peut également représenter les erreurs de mesure lorsque les coefficients de la matrice A sont obtenus expérimentalement. Nous noterons encore $\vec{x}^*$, la solution du système perturbé. On a donc la relation:

$$\vec{x} = A^{-1}\vec{b} = A^{-1}((A + E)\vec{x}^*) = (I + A^{-1}E)\vec{x}^* = \vec{x}^* + A^{-1}E\vec{x}^*$$

Ou en conclut que:

$$\vec{x} - \vec{x}^* = A^{-1}E\vec{x}^*$$

Donc, en vertu des relations 3.36 et 3.37:

$$||\vec{x} - \vec{x}^*|| \leq ||A^{-1}||\,||E||\,||\vec{x}^*|| = \frac{||A||\,||A^{-1}||\,||E||\,||\vec{x}^*||}{||A||}$$

d'où l'on tire le théorème suivant.

Théorème 3.5

$$\frac{||\vec{x} - \vec{x}^*||}{||\vec{x}^*||} \leq \text{cond}A\,\frac{||E||}{||A||} \qquad \square \qquad (3.46)$$

Remarque 3.31
Les remarques suivantes permettent de bien mesurer la portée de l'inégalité 3.46.

1. Le terme de gauche est une approximation de l'erreur relative entre la solution exacte et la solution du système perturbé. (On devrait avoir $||\vec{x}||$ au dénominateur pour représenter vraiment l'erreur relative.)

2. Le terme de droite est en quelque sorte l'erreur relative liée aux coefficients de la matrice A multipliée par le conditionnement de A.
3. Si cond$A$ est petit, une petite perturbation sur la matrice A entraîne un petite perturbation sur la solution $\vec{x}$.
4. Par contre, si cond$A$ est grand, une petite perturbation sur la matrice A pourrait résulter en une très grande perturbation sur la solution du système. Il est par conséquent possible que les résultats numériques soient peu précis et même, dans certains cas, complètement faux. □

Remarque 3.32
Très souvent, la perturbation E de la matrice A provient des erreurs dues à la représentation des nombres sur ordinateur. Par définition de la précision machine ϵ et de la norme l_∞, on a dans ce cas:

$$||E||_\infty \leq \epsilon \; ||A||_\infty$$

ce qui permet de réécrire la conclusion 3.46 du théorème sous la forme:

$$\frac{||\vec{x} - \vec{x}^*||_\infty}{||\vec{x}^*||_\infty} \leq \epsilon \; \text{cond}A = \epsilon \; ||A||_\infty \; ||A^{-1}||_\infty \tag{3.47}$$

On constate que, plus le conditionnement est élevé, plus la précision machine ϵ doit être petite. Si la simple précision est insuffisante, on recourt à la double précision. □

Exemple 3.23
La matrice:

$$A = \begin{bmatrix} 1{,}012 & -2{,}132 & 3{,}104 \\ -2{,}132 & 4{,}096 & -7{,}013 \\ 3{,}014 & -7{,}013 & 0{,}014 \end{bmatrix}$$

a comme inverse:

$$A^{-1} = \begin{bmatrix} -13{,}729 & -6{,}0755 & 0{,}625\,40 \\ -6{,}0755 & -2{,}6888 & 0{,}133\,99 \\ 0{,}625\,40 & 0{,}133\,99 & -0{,}111\,87 \end{bmatrix}$$

On a alors $||A||_\infty = 13{,}241$ et $||A^{-1}||_\infty = 20{,}43$. On conclut que le conditionnement de la matrice A est:

$$\text{cond}A = (13{,}241)(20{,}43) = 270{,}51$$

• • • • •

Remarque 3.33
En utilisant une autre norme matricielle, on obtiendrait un conditionnement différent. Toutefois, on pourrait montrer que le conditionnement, s'il est grand dans une norme, sera grand dans toutes les normes. □

Exemple 3.24
La matrice:

$$A = \begin{bmatrix} 3{,}02 & -1{,}05 & 2{,}53 \\ 4{,}33 & 0{,}56 & -1{,}78 \\ -0{,}83 & -0{,}54 & 1{,}47 \end{bmatrix}$$

a comme inverse:

$$A^{-1} = \begin{bmatrix} 5{,}661 & -7{,}273 & -18{,}55 \\ 200{,}5 & -268{,}3 & -669{,}9 \\ 76{,}85 & -102{,}6 & -255{,}9 \end{bmatrix}$$

Pour cette matrice, $||A||_\infty = 6{,}67$ et $||A^{-1}||_\infty = 1138{,}7$. Le conditionnement de la matrice est donc 7595, ce qui est le signe d'une matrice mal conditionnée.

• • • • •

Exemple 3.25
Nous avons déjà considéré la matrice:

$$A = \begin{bmatrix} 0{,}0003 & 3{,}0 \\ 1{,}0 & 1{,}0 \end{bmatrix}$$

dont l'inverse est:

$$A^{-1} = \begin{bmatrix} -0{,}333\,367 & 1{,}0001 \\ 0{,}333\,367 & 1{,}0001 \times 10^{-4} \end{bmatrix}$$

On a ainsi un conditionnement d'environ 4, ce qui est relativement faible. Nous avons vu que la résolution d'un système linéaire à l'aide de cette matrice, sans effectuer de permutation de lignes, aboutit à de mauvais résultats. Cela démontre bien qu'un algorithme mal choisi (la décomposition LU sans permutation de lignes dans ce cas) peut s'avérer inefficace, et ce même si la matrice est bien conditionnée.

●●●●●

3.8 Systèmes non linéaires

Les phénomènes non linéaires sont extrêmement courants en pratique. Ils sout sans doute plus fréquents que les phénomènes linéaires. Dans cette section, nous examinons les systèmes non linéaires et nous montrons comment les résoudre à l'aide d'une suite de problèmes linéaires, auxquels on peut appliquer diverses techniques de résolution comme la décomposition LU.

Le problème consiste à trouver le ou les vecteurs $\vec{x} = [x_1 \ x_2 \ x_3 \ \cdots \ x_n]^T$ vérifiant les n équations non linéaires suivantes:

$$\begin{array}{rcl} f_1(x_1, x_2, x_3, \cdots, x_n) & = & 0 \\ f_2(x_1, x_2, x_3, \cdots, x_n) & = & 0 \\ f_3(x_1, x_2, x_3, \cdots, x_n) & = & 0 \\ \vdots & & \vdots \\ f_n(x_1, x_2, x_3, \cdots, x_n) & = & 0 \end{array} \qquad (3.48)$$

où les f_i sont des fonctions de n variables que nous supposons différentiables. Contrairement aux systèmes linéaires, il n'y a pas de condition simple associée aux systèmes non linéaires qui permette d'assurer l'existence et l'unicité de la solution. Le plus souvent, il existe plusieurs solutions possibles et seul le contexte indique laquelle est la bonne.

Les méthodes de résolution des systèmes non linéaires sont nombreuses. Notamment, presque toutes les méthodes du chapitre 2 peuvent être généralisées aux systèmes non linéaires. Pour éviter de surcharger notre exposé, nous ne présentons que la méthode la plus importante et la plus utilisée en pratique, soit la *méthode de Newton*.

L'application de cette méthode à un système de deux équations non linéaires est suffisante pour illustrer le cas général. Il serait également bon de réviser le développement de la méthode de Newton pour une équation non linéaire (voir le chapitre 2) puisque le raisonnement est le même pour les systèmes.

Considérons donc le système:

$$\begin{aligned} f_1(x_1, x_2) &= 0 \\ f_2(x_1, x_2) &= 0 \end{aligned}$$

Soit (x_1^0, x_2^0), une approximation initiale de la solution de ce système. *Cette approximation initiale est cruciale et doit toujours être choisie avec soin.* Le but de ce qui suit est de déterminer une correction $(\delta x_1, \delta x_2)$ à (x_1^0, x_2^0) de telle sorte que:

$$\begin{aligned} f_1(x_1^0 + \delta x_1, x_2^0 + \delta x_2) &= 0 \\ f_2(x_1^0 + \delta x_1, x_2^0 + \delta x_2) &= 0 \end{aligned}$$

Pour déterminer $(\delta x_1, \delta x_2)$, il suffit maintenant de faire un développement de Taylor en deux variables pour chacune des deux fonctions (voir la section 1.6.2; voir aussi Thomas et Finney, réf. [22]):

$$\begin{aligned} 0 &= f_1(x_1^0, x_2^0) + \frac{\partial f_1}{\partial x_1}(x_1^0, x_2^0)\, \delta x_1 + \frac{\partial f_1}{\partial x_2}(x_1^0, x_2^0)\, \delta x_2 + \cdots \\ 0 &= f_2(x_1^0, x_2^0) + \frac{\partial f_2}{\partial x_1}(x_1^0, x_2^0)\, \delta x_1 + \frac{\partial f_2}{\partial x_2}(x_1^0, x_2^0)\, \delta x_2 + \cdots \end{aligned}$$

Dans les relations précédentes, les pointillés désignent des termes d'ordre supérieur ou égal à deux et faisant intervenir les dérivées partielles d'ordre correspondant. Pour déterminer $(\delta x_1, \delta x_2)$, il suffit de négliger les termes d'ordre supérieur et d'écrire:

$$\begin{aligned} \frac{\partial f_1}{\partial x_1}(x_1^0, x_2^0)\, \delta x_1 + \frac{\partial f_1}{\partial x_2}(x_1^0, x_2^0)\, \delta x_2 &= -f_1(x_1^0, x_2^0) \\ \frac{\partial f_2}{\partial x_1}(x_1^0, x_2^0)\, \delta x_1 + \frac{\partial f_2}{\partial x_2}(x_1^0, x_2^0)\, \delta x_2 &= -f_2(x_1^0, x_2^0) \end{aligned}$$

ou encore sous forme matricielle:

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x_1^0, x_2^0) & \frac{\partial f_1}{\partial x_2}(x_1^0, x_2^0) \\ \frac{\partial f_2}{\partial x_1}(x_1^0, x_2^0) & \frac{\partial f_2}{\partial x_2}(x_1^0, x_2^0) \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} f_1(x_1^0, x_2^0) \\ f_2(x_1^0, x_2^0) \end{bmatrix}$$

Ce système *linéaire* s'écrit également sous une forme plus compacte:

$$J(x_1^0, x_2^0)\, \vec{\delta x} = -\vec{R}(x_1^0, x_2^0) \tag{3.49}$$

où $J(x_1^0, x_2^0)$ désigne la matrice des dérivées partielles ou *matrice jacobienne* évaluée au point (x_1^0, x_2^0), où $\vec{\delta x}$ est le vecteur des corrections relatives à chaque variable et où $-\vec{R}(x_1^0, x_2^0)$ est le *vecteur résidu* évalué en (x_1^0, x_2^0). Le déterminant de la matrice jacobienne est appelé le *jacobien*. Le jacobien doit bien entendu être différent de 0 pour que la matrice jacobienne soit inversible. On pose ensuite:

$$\begin{array}{rcl} x_1^1 & = & x_1^0 + \delta x_1 \\ x_2^1 & = & x_2^0 + \delta x_2 \end{array}$$

qui est la nouvelle approximation de la solution du système non linéaire. On cherchera par la suite à corriger (x_1^1, x_2^1) d'une nouvelle quantité $(\vec{\delta x})$, et ce jusqu'à la convergence.

De manière plus générale, on pose:

$$J(\vec{x}^i) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1}(\vec{x}^i) & \dfrac{\partial f_1}{\partial x_2}(\vec{x}^i) & \cdots & \dfrac{\partial f_1}{\partial x_n}(\vec{x}^i) \\ \dfrac{\partial f_2}{\partial x_1}(\vec{x}^i) & \dfrac{\partial f_2}{\partial x_2}(\vec{x}^i) & \cdots & \dfrac{\partial f_2}{\partial x_n}(\vec{x}^i) \\ \vdots & \vdots & \ddots & \\ \dfrac{\partial f_n}{\partial x_1}(\vec{x}^i) & \dfrac{\partial f_n}{\partial x_2}(\vec{x}^i) & \cdots & \dfrac{\partial f_n}{\partial x_n}(\vec{x}^i) \end{bmatrix}$$

c'est-à-dire la matrice jacobienne évaluée au point $\vec{x}^i = (x_1^i, x_2^i, \cdots x_n^i)$. De plus on pose:

$$\vec{R}(\vec{x}^i) = \begin{bmatrix} f_1(\vec{x}^i) \\ f_2(\vec{x}^i) \\ \vdots \\ f_n(\vec{x}^i) \end{bmatrix} \qquad \vec{\delta x} = \begin{bmatrix} \delta x_1 \\ \delta x_2 \\ \vdots \\ \delta x_n \end{bmatrix}$$

pour en arriver à l'algorithme général suivant.

Algorithme 3.2: *Méthode de Newton appliquée aux systèmes*

1. Étant donné ϵ, un critère d'arrêt
2. Étant donné N, le nombre maximal d'itérations
3. Étant donné $\vec{x}^0 = [x_1^0 \ \ x_2^0 \ \ \cdots \ \ x_n^0]^T$, une approximation initiale de la solution du système
4. Résoudre le système linéaire:

$$J(\vec{x}^i)\, \vec{\delta x} = -\vec{R}(\vec{x}^i) \tag{3.50}$$

et poser:

$$\vec{x}^{i+1} = \vec{x}^i + \vec{\delta x}$$

5. Si $\dfrac{||\vec{\delta x}||}{||\vec{x_{i+1}}||} < \epsilon$ et $||\vec{R}(\vec{x}^{i+1})|| \leq \epsilon$:
   - convergence atteinte
   - écrire la solution $\vec{x}^{i+1}$
   - arrêt
6. Si le nombre maximal d'itérations N est atteint:
   - convergence non atteinte en N itérations
   - arrêt
7. Retour à l'étape 4 □

Exemple 3.26

On cherche à trouver l'intersection de la courbe $x_2 = e^{x_1}$ et du cercle de rayon 4 centré à l'origine d'équation $x_1^2 + x_2^2 = 16$. L'intersection de ces courbes est une solution de:

$$\begin{aligned} e^{x_1} - x_2 &= 0 \\ x_1^2 + x_2^2 - 16 &= 0 \end{aligned}$$

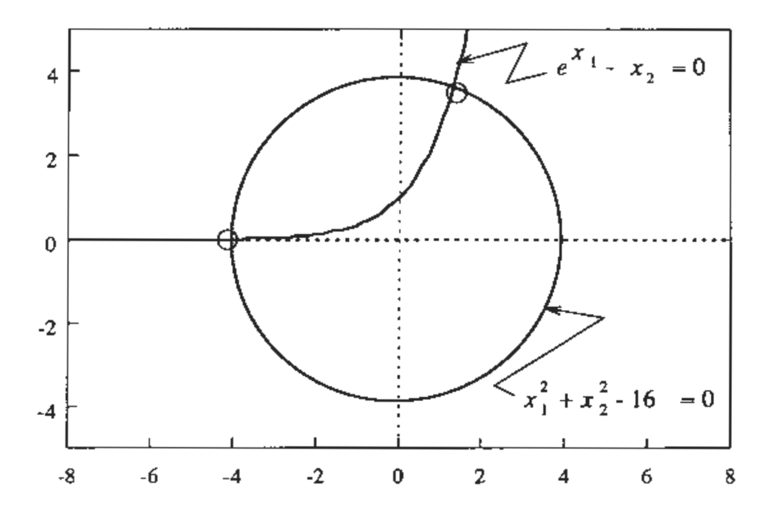


Figure 3.1: Intersection de deux courbes

La première étape consiste à calculer la matrice jacobienne de dimension 2. Dans ce cas, on a:

$$J(x_1, x_2) = \begin{bmatrix} e^{x_1} & -1 \\ 2x_1 & 2x_2 \end{bmatrix}$$

Un graphique de ces deux courbes montre qu'il y a deux solutions à ce problème non linéaire (voir la figure 3.1). La première solution se trouve près du point $(-4\,,\,0)$ et la deuxième, près de $(2{,}8\,,\,2{,}8)$. Prenons le point $(2{,}8\,,\,2{,}8)$ comme approximation initiale de la solution de ce système non linéaire, c'est-à-dire $\vec{x}_0 = [2{,}8\ \ 2{,}8]^T$.

1. Itération 1:

   Le système 3.50 devient:

$$\begin{bmatrix} e^{2,8} & -1 \\ 2(2{,}8) & 2(2{,}8) \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} e^{2,8} - 2{,}8 \\ (2{,}8)^2 + (2{,}8)^2 - 16 \end{bmatrix}$$

   c'est-à-dire

$$\begin{bmatrix} 16{,}445 & -1 \\ 5{,}6 & 5{,}6 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} 13{,}645 \\ -0{,}3200 \end{bmatrix}$$

   dont la solution est $\vec{\delta x} = [-0{,}778\,90\ \ 0{,}836\,04]^T$. La nouvelle approximation de la solution est donc:

$$\begin{array}{rcccccc} x_1^1 & = & x_1^0 + \delta x_1 & = & 2{,}8 - 0{,}778\,90 & = & 2{,}0211 \\ x_2^1 & = & x_2^0 + \delta x_2 & = & 2{,}8 + 0{,}836\,04 & = & 3{,}636\,04 \end{array}$$

2. Itération 2:

   On effectue une deuxième itération à partir de (x_1^1, x_2^1). Le système 3.50 devient alors:

$$\begin{bmatrix} e^{2,0211} & -1 \\ 2(2,0211) & 2(3,636\,04) \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} e^{2,0211} - 3,636\,04 \\ (2,0211)^2 + (3,636\,04)^2 - 16 \end{bmatrix}$$

   c'est-à-dire

$$\begin{bmatrix} 7,5466 & -1 \\ 4,0422 & 7,2721 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} 3,9106 \\ 1,3056 \end{bmatrix}$$

   dont la solution est $\vec{\delta x} = [-0,5048 \;\; 0,101\,06]^T$. On a maintenant:

$$\begin{array}{lllllll} x_1^2 & = & x_1^1 + \delta x_1 & = & 2,0211 - 0,504\,80 & = & 1,5163 \\ x_2^2 & = & x_2^1 + \delta x_2 & = & 3,636\,04 + 0,101\,06 & = & 3,7371 \end{array}$$

3. Itération 3:

   À la troisième itération, on doit résoudre:

$$\begin{bmatrix} 4,5554 & -1 \\ 3,0326 & 7,4742 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} 0,818\,24 \\ 0,265\,08 \end{bmatrix}$$

   ce qui entraîne que $\vec{\delta x} = [-0,172\,08 \;\; 0,034\,355]^T$. La nouvelle solution est:

$$\begin{array}{lllllll} x_1^3 & = & x_1^2 + \delta x_1 & = & 1,5163 - 0,172\,08 & = & 1,3442 \\ x_2^3 & = & x_2^2 + \delta x_2 & = & 3,7371 + 0,034\,355 & = & 3,7715 \end{array}$$

4. Itération 4:

   Le système linéaire à résoudre est:

$$\begin{bmatrix} 3,8351 & -1 \\ 2,6884 & 7,5430 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} 0,063\,617 \\ 0,031\,086 \end{bmatrix}$$

   ce qui entraîne que $\vec{\delta x} = [-0,016\,1616 \;\; 0,016\,3847]^T$. La nouvelle approximation de la solution est:

$$\begin{array}{lllllll} x_1^4 & = & x_1^3 + \delta x_1 & = & 1,3442 - 0,016\,1616 & = & 1,3280 \\ x_2^4 & = & x_2^3 + \delta x_2 & = & 3,7715 + 0,016\,3847 & = & 3,7731 \end{array}$$

5. Itération 5:

   Enfin, à partir de $[1{,}3280 \quad 3{,}7731]^T$, on doit résoudre:

$$\begin{bmatrix} 3{,}7735 & -1 \\ 2{,}6560 & 7{,}5463 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} 0{,}348\,86 \times 10^{-3} \\ 0{,}169\,46 \times 10^{-3} \end{bmatrix}$$

   dont la solution est $\vec{\delta x} = [9{,}03 \times 10^{-5} \quad 9{,}25 \times 10^{-6}]^T$. La solution du système non linéaire devient:

$$\begin{array}{rclcl} x_1^5 & = & x_1^4 + \delta x_1 & = & 1{,}3281 \\ x_2^5 & = & x_2^4 + \delta x_2 & = & 3{,}7731 \end{array}$$

On déduit la convergence de l'algorithme de Newton du fait que les modules de $\vec{\delta x}$ et de $\vec{R}$ diminuent avec les itérations.

●●●●●

Remarque 3.34

1. La convergence de la méthode de Newton dépend de l'approximation initiale $\vec{x}^0$ de la solution. *Un mauvais choix de $\vec{x}^0$ peut résulter en un algorithme divergent.*

2. On peut démontrer que, lorsqu'il y a convergence de l'algorithme, cette convergence est généralement quadratique dans le sens suivant:

$$||\vec{x} - \vec{x}^{i+1}|| \simeq C\, ||\vec{x} - \vec{x}^i||^2 \qquad (3.51)$$

   Cela signifie que la norme de l'erreur à l'itération $i+1$ est approximativement égale à une constante C multipliée par le carré de la norme de l'erreur à l'étape i. *L'analogie est évidente avec le cas d'une seule équation non linéaire étudié au chapitre 2.*

3. La convergence quadratique est perdue si la matrice jacobienne est singulière au point $\vec{x}$, solution du système non linéaire. *Encore une fois, ce comportement est analogue au cas d'une seule équation où la méthode de Newton perd sa convergence quadratique si la racine est de multiplicité plus grande que 1 ($f'(r) = 0$).*

4. Pour obtenir la convergence quadratique, on doit calculer et décomposer une matrice de taille n sur n à chaque itération. De plus, il faut fournir à un éventuel programme informatique les n fonctions $f_i(\vec{x})$ et les n^2 dérivées partielles de ces fonctions. Cela peut devenir rapidement fastidieux et coûteux lorsque la dimension n du système est grande.

5. Il existe une variante de la méthode de Newton qui évite le calcul des n^2 dérivées partielles et qui ne nécessite que les n fonctions $f_i(\vec{x})$. La *méthode de Newton modifiée* consiste à remplacer les dérivées partielles par des différences centrées (voir le chapitre 6). On utilise alors l'approximation du second ordre ($O(h^2)$):

$$\frac{\partial f_i}{\partial x_j}(x_1, x_2, \cdots, x_n) \simeq$$

$$\frac{f_i(x_1, \cdots, x_{j-1}, x_j + h, \cdots, x_n) - f_i(x_1, \cdots, x_{j-1}, x_j - h, \cdots, x_n)}{2h} \tag{3.52}$$

Cette approximation introduit une petite erreur dans le calcul de la matrice jacobienne, mais généralement la convergence est quand même très rapide. □

Exemple 3.27
Dans l'exemple précédent, le calcul du premier terme de la matrice jacobienne de la première itération donnait:

$$\frac{\partial f_1}{\partial x_1}(2{,}8\ ,\ 2{,}8) = e^{2{,}8} = 16{,}444\,646\,77$$

tandis que l'approximation 3.52 donne pour $h = 0{,}001$:

$$\frac{f_1(2{,}801\ ,\ 2{,}8) - f_1(2{,}799\ ,\ 2{,}8)}{(2)(0{,}001)} = 16{,}444\,65$$

On constate que l'erreur introduite est minime.

● ● ● ● ●

3.9 Applications

3.9.1 Calcul des tensions dans une ferme

Une *ferme* est une structure bidimensionnelle relativement simple composée de *membrures* métalliques droites jointes par des *rotules*. Les équations

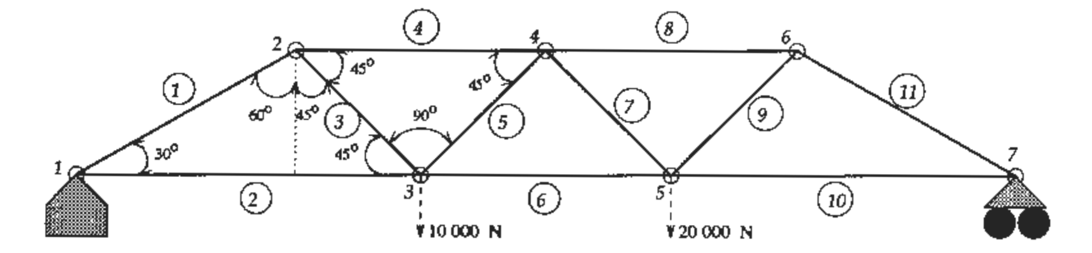


Figure 3.2: Membrures et rotules d'une ferme

de la mécanique nous assurent que les efforts dans une telle structure se réduisent à des tractions-compressions. Cette structure, dont on néglige le poids, est par la suite soumise à une charge qui provoque des tensions ou des compressions dans les membrures. Cela provoque également des déplacements de faible amplitude des rotules. Typiquement, on a la situation de la figure 3.2, où une ferme représentant grossièrement un pont de chemin de fer est constituée de 11 membrures et de 7 rotules. Le première rotule est fixée sur un support qui prévient tout déplacement et qui exerce une poussée horizontale H_1 et une poussée verticale V_1. À l'autre extrémité, la rotule 7 est fixée sur un rail qui ne prévient que les déplacements verticaux et qui fournit une poussée verticale V_2. Des charges de $-10\,000$ N et de $-20\,000$ N sont exercées sur les rotules 3 et 5 respectivement. Les forces sont considérées positives si elles agissent vers le haut suivant la verticale et vers la droite suivant l'horizontale.

Pour calculer les tensions T_i dans les membrures, il suffit de recourir aux équations d'équilibre horizontal et vertical à chaque rotule. *Nous supposons que chaque membrure est en tension. Si nous obtenons un résultat négatif pour l'un des T_i, cela signifie que la membrure en question est en compression.* Une membrure en tension exerce une poussée sur les rotules qui, en réaction, poussent sur les membrures. L'équilibre est atteint lorsque ces poussées sont compensées par les charges externes agissant sur la rotule. La figure 3.3 illustre les forces exercées sur chaque rotule. À l'aide de cette figure, on peut établir les conditions d'équilibre.

- Rotule 1: Fixée
- Rotule 2: Équilibre horizontal et vertical

$$\begin{aligned}
-T_1 \cos 30^\circ + T_3 \cos 45^\circ + T_4 &= 0 \\
-T_1 \sin 30^\circ - T_3 \sin 45^\circ &= 0
\end{aligned}$$

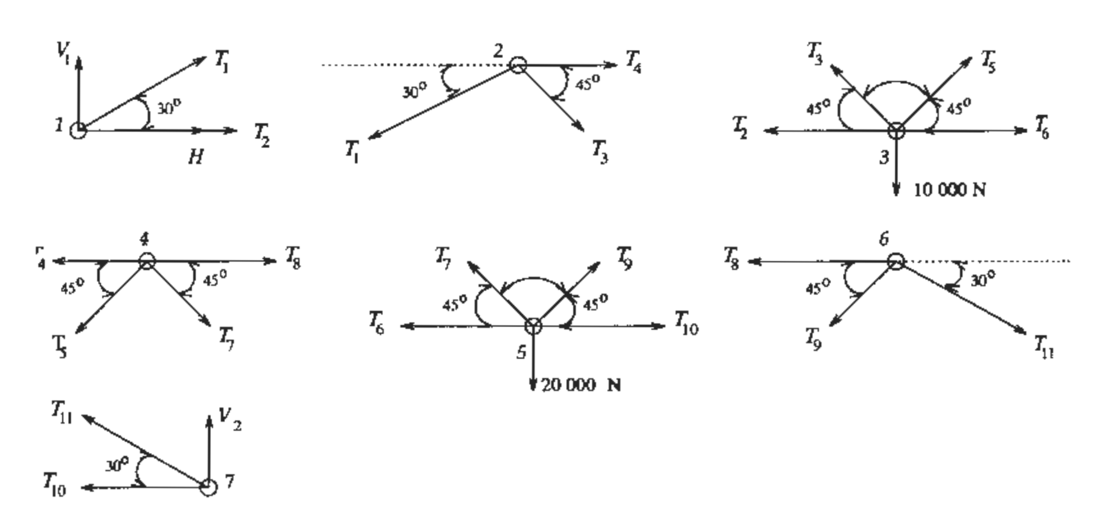


Figure 3.3: Forces exercées sur les rotules

- Rotule 3: Équilibre horizontal et vertical

$$-T_2 - T_3 \cos 45^\circ + T_5 \cos 45^\circ + T_6 = 0$$

$$T_3 \sin 45^\circ + T_5 \sin 45^\circ - 10\,000 = 0$$

- Rotule 4: Équilibre horizontal et vertical

$$-T_4 - T_5 \cos 45^\circ + T_7 \cos 45^\circ + T_8 = 0$$

$$-T_5 \sin 45^\circ - T_7 \sin 45^\circ = 0$$

- Rotule 5: Équilibre horizontal et vertical

$$-T_6 - T_7 \cos 45^\circ + T_9 \cos 45^\circ + T_{10} = 0$$

$$T_7 \sin 45^\circ + T_9 \sin 45^\circ - 20\,000 = 0$$

- Rotule 6: Équilibre horizontal et vertical

$$-T_8 - T_9 \cos 45^\circ + T_{11} \cos 30^\circ = 0$$

$$-T_9 \sin 45^\circ - T_{11} \sin 30^\circ = 0$$

- Rotule 7: Équilibre horizontal seulement

$$-T_{10} - T_{11} \cos 30^\circ = 0$$

Sous forme matricielle, on obtient un système linéaire de 11 équations de la forme:

$$A\vec{T} = \vec{b}$$

La matrice A complète est:

$$A = \begin{bmatrix} -a & 0 & b & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0{,}5 & 0 & -b & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & -b & 0 & b & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & b & 0 & b & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -b & 0 & b & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -b & 0 & -b & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & -b & 0 & b & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & b & 0 & b & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -b & 0 & a \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -b & 0 & -0{,}5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -a \end{bmatrix}$$

où $a = \cos 30^\circ = 0{,}866\,025\,404$ et $b = \cos 45^\circ = \sin 45^\circ = 0{,}707\,106\,781$.

$$\vec{T} = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \\ T_6 \\ T_7 \\ T_8 \\ T_9 \\ T_{10} \\ T_{11} \end{bmatrix} \qquad \vec{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 10\,000 \\ 0 \\ 0 \\ 0 \\ 20\,000 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Une décomposition LU permet la résolution de ce système. On obtient les valeurs suivantes pour les tensions (en newtons) exercées sur les différentes

membrures:

$$\vec{T} = \begin{bmatrix} -27\,320{,}5 \\ 23\,659{,}5 \\ 19\,321{,}4 \\ -37\,319{,}7 \\ -5177{,}1 \\ 40\,980{,}0 \\ 5177{,}1 \\ -44\,640{,}2 \\ 23\,111{,}4 \\ 28\,300{,}5 \\ -32\,679{,}5 \end{bmatrix}$$

Ainsi, les membrures 1, 4, 5, 8 et 11 sont en compression et toutes les autres sont en tension. Pour vérifier la validité de ces résultats, on peut calculer les forces verticales V_1 et V_2. L'équilibre des forces verticales à la rotule 1 nous assure que:

$$V_1 = -T_1 \sin 30^{\circ} = 13\,660{,}25 \text{ N}$$

et de la même manière à la rotule 7:

$$V_2 = -T_{11} \sin 30^{\circ} = 16\,339{,}75 \text{ N}$$

Le total de $V_1 + V_2$ est bien 30 000 N, correspondant à la charge totale.

On pourrait maintenant facilement étudier l'effet de différentes répartitions de charge sur cette structure. Seul le vecteur $\vec{b}$ serait affecté. La matrice A ne change pas tant que l'on ne modifie pas la ferme elle-même.

3.9.2 Deuxième modèle de viscosité

Au chapitre précédent (voir la section 2.6.2), nous avons considéré un modèle de viscosité exprimé par la loi puissance de la forme:

$$\eta = \eta_0 \dot{\gamma}^{\beta-1}$$

Un modèle plus réaliste est le *modèle de Carreau*[1] qui est de la forme:

$$\eta = \eta_0 (1 + \lambda^2 \dot{\gamma}^2)^{\frac{\beta-1}{2}} \tag{3.53}$$

[1] P. Carreau est professeur au Département de génie chimique de l'École Polytechnique de Montréal.

Les paramètres η_0, λ et β sont déterminés en fonction des mesures expérimentales. Nous sommes donc amenés à rechercher les valeurs minimales de la fonction de moindres carrés:

$$F(\eta_0, \lambda, \beta) = \frac{1}{2}\sum_{i=1}^{npt}(\eta_0(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-1}{2}} - \eta_i)^2$$

où npt est le nombre de données mesurées à l'aide d'un rhéomètre. Nous utilisons les mêmes données rhéologiques qu'à la section 2.6.2, ce qui nous permettra de comparer le modèle de Carreau et la loi puissance. Le minimum sera atteint lorsque:

$$\frac{\partial F(\eta_0, \lambda, \beta)}{\partial \eta_0} = \frac{\partial F(\eta_0, \lambda, \beta)}{\partial \lambda} = \frac{\partial F(\eta_0, \lambda, \beta)}{\partial \beta} = 0$$

c'est-à-dire lorsque[2]:

$$\begin{aligned}
\frac{\partial F(\eta_0, \lambda, \beta)}{\partial \eta_0} &= \sum_{i=1}^{npt}(\eta_0(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-1}{2}} - \eta_i)(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-1}{2}} = 0 \\
\frac{\partial F(\eta_0, \lambda, \beta)}{\partial \lambda} &= \sum_{i=1}^{npt}(\eta_0(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-1}{2}} - \eta_i)\eta_0\frac{(\beta-1)}{2}(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-3}{2}}2\lambda\dot{\gamma}_i^2 = 0 \\
\frac{\partial F(\eta_0, \lambda, \beta)}{\partial \beta} &= \sum_{i=1}^{npt}(\eta_0(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-1}{2}} - \eta_i)\frac{\eta_0}{2}(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-1}{2}}\ln(1 + \lambda^2\dot{\gamma}_i^2) = 0
\end{aligned}$$

que l'on peut simplifier légèrement en extrayant de la sommation les termes qui ne dépendent pas de l'indice i. On obtient en bout de course le système de trois équations non linéaires suivant:

$$\begin{aligned}
\sum_{i=1}^{npt}(\eta_0(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-1}{2}} - \eta_i)(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-1}{2}} &= 0 \\
\sum_{i=1}^{npt}(\eta_0(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-1}{2}} - \eta_i)(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-3}{2}}\dot{\gamma}_i^2 &= 0 \qquad (3.54) \\
\sum_{i=1}^{npt}(\eta_0(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-1}{2}} - \eta_i)(1 + \lambda^2\dot{\gamma}_i^2)^{\frac{\beta-1}{2}}\ln(1 + \lambda^2\dot{\gamma}_i^2) &= 0
\end{aligned}$$

[2]La dérivée de $a^{f(x)}$ est $a^{f(x)}f'(x)\ln a$.

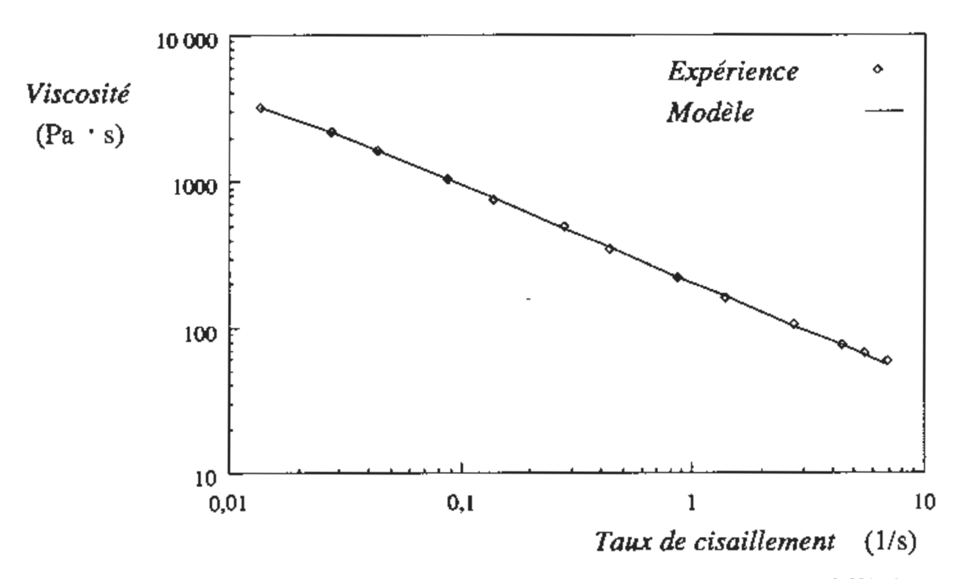


Figure 3.4: Loi de Carreau: $\eta = 4926{,}08(1 + (116{,}922)^2\dot{\gamma}^2)^{\frac{0{,}331-1}{2}}$

On a utilisé la méthode de Newton modifiée (voir l'équation 3.52) pour résoudre ce système. À partir de l'approximation initiale $[5200 \;\; 140 \;\; 0{,}38]^T$, la méthode a convergé en 8 itérations vers la solution $[4926{,}08 \;\; 116{,}922 \;\; 0{,}331]^T$. Une comparaison du modèle de Carreau avec les données expérimentales est présentée à la figure 3.4.

On remarque une meilleure correspondance entre les valeurs expérimentales et celles calculées à l'aide du modèle de Carreau que celle qui a été obtenue par la loi puissance (voir la figure 2.9).

3.10 Exercices

1. Soit la matrice:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

   Identifier les matrices W qui permettent d'effectuer les opérations suivantes:

   a) $\vec{l}_2 \leftarrow \vec{l}_2 - 3\vec{l}_1$

   b) $\vec{l}_2 \leftrightarrow \vec{l}_3$

   c) $\vec{l}_2 \leftarrow 5\vec{l}_2$

   d) $\vec{l}_3 \leftarrow \vec{l}_3 + 5\vec{l}_2$

   Calculer le déterminant de chaque matrice W et son inverse W^{-1}.

2. Résoudre les systèmes triangulaires suivants:

   a)

$$\begin{bmatrix} 3 & 0 & 0 \\ 1 & 5 & 0 \\ 2 & 4 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 13 \\ 20 \end{bmatrix}$$

   b)

$$\begin{bmatrix} 1 & 3 & 4 \\ 0 & 3 & 5 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ -4 \\ 6 \end{bmatrix}$$

   Calculer le déterminant de ces deux matrices.

3. En indiquant bien les opérations effectuées sur les lignes, utiliser l'élimination de Gauss pour triangulariser les systèmes linéaires suivants:

   a)

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 2 & 3 \\ -1 & -3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \\ 2 \end{bmatrix}$$

   b)

$$\begin{bmatrix} 1 & 2 & 1 & 4 \\ 2 & 0 & 4 & 3 \\ 4 & 2 & 2 & 1 \\ -3 & 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 13 \\ 28 \\ 20 \\ 6 \end{bmatrix}$$

Calculer le déterminant de ces deux matrices.

4. Obtenir les matrices W correspondant aux opérations élémentaires effectuées sur les matrices de l'exercice précédent. Montrer pour ces deux exemples que la méthode d'élimination de Gauss est équivalente à une décomposition LU.

5. a) Résoudre le système linéaire suivant par élimination de Gauss et en utilisant l'arithmétique flottante à 4 chiffres, mais **sans permutation de lignes**.

$$\begin{bmatrix} 0{,}729 & 0{,}81 & 0{,}9 \\ 1 & 1 & 1 \\ 1{,}331 & 1{,}21 & 1{,}1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0{,}6867 \\ 0{,}8338 \\ 1 \end{bmatrix}$$

   b) Résoudre le même système linéaire en arithmétique flottante à 4 chiffres, mais cette fois en permutant les lignes de façon à avoir le plus grand pivot possible.

   c) Comparer les deux solutions numériques avec la solution exacte $\vec{x} = [0{,}2245\ 0{,}2814\ 0{,}3279]^T$ et calculer les erreurs relatives en norme l_∞.

6. On veut résoudre le système linéaire suivant par élimination de Gauss.

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

   a) La matrice de ce système linéaire est-elle singulière?

   b) Combien de solutions ce système linéaire possède-t-il?

7. a) Effectuer l'élimination de Gauss (sans permutation de lignes) sur le système:

$$\begin{bmatrix} 2 & -6\alpha \\ 3\alpha & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ \beta \end{bmatrix}$$

   b) Calculer le déterminant de A en vous servant de l'élimination de Gauss.

   c) Déterminer les valeurs de α et de β pour lesquelles la matrice A est non inversible (singulière).

   d) Que pouvez-vous dire de la solution de ce système quand $\alpha = 1/3$ et $\beta = 1$?

8. Résoudre les systèmes linéaires suivants par la méthode de décomposition LU de Crout (sans permutation de lignes).

   a)
$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 2 & 3 \\ -1 & -3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \\ 2 \end{bmatrix}$$

   b)
$$\begin{bmatrix} 1 & 2 & 1 & 4 \\ 2 & 0 & 4 & 3 \\ 4 & 2 & 2 & 1 \\ -3 & 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 13 \\ 28 \\ 20 \\ 6 \end{bmatrix}$$

9. Résoudre le système linéaire:
$$\begin{bmatrix} 1 & 2 & 6 \\ 4 & 8 & -1 \\ -2 & 3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 23 \\ 17 \\ 10 \end{bmatrix}$$
   par décomposition LU avec permutation de lignes.

10. La matrice:
$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 7 & 18 \\ 4 & 13 & 38 \end{bmatrix}$$
   possède la décomposition LU suivante (notation compacte, obtenue sans permutation de lignes):
$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 4 & 5 & 6 \end{bmatrix}$$
   En utilisant la méthode de Crout, on a résolu les systèmes $A\vec{x} = \vec{b}$ suivants:

   - 1) Si $\vec{b} = (1 \ \ 0 \ \ 0)^T$:
$$\vec{x} = (1{,}7777 \quad -0{,}222\,22 \quad -0{,}111\,11)^T$$
   - 2) Si $\vec{b} = (0 \ \ 1 \ \ 0)^T$:
$$\vec{x} = (-2{,}0555 \quad 1{,}4444 \quad -0{,}277\,77)^T$$

En complétant au besoin les données précédentes, répondre aux questions suivantes:

a) dét $A =$

b) $||A||_\infty =$

c) $A^{-1} =$

d) cond$A =$ (Utiliser $|| \; ||_\infty$.)

11. La matrice:

$$\begin{bmatrix} 2 & -1 & 0 \\ 4 & -1 & 2 \\ -6 & 2 & 0 \end{bmatrix}$$

possède la décomposition LU suivante (notation compacte, obtenue sans permutation de lignes):

$$\begin{bmatrix} 2 & -1/2 & 0 \\ 4 & 1 & 2 \\ -6 & -1 & 2 \end{bmatrix}$$

En utilisant la décomposition LU, effectuer les opérations suivantes:

a) Calculer dét A.

b) Résoudre le système linéaire $A\vec{x} = \vec{b}$ où:

$$\vec{b} = \begin{bmatrix} -2 \\ 14 \\ 12 \end{bmatrix}$$

c) **Sans calculer** A^2, résoudre le système $A^2\vec{x} = \vec{b}$ pour $\vec{b}$ donné en b). (Rappel: $A^2\vec{x} = A(A\vec{x})$.)

12. Résoudre le système linéaire:

$$\begin{bmatrix} \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 8 \\ 8 \end{bmatrix}$$

en arithmétique flottante à 3 chiffres avec pivotage. Comparer le résultat numérique avec la solution exacte:

$$\vec{x} = (-2952/13 \;\; 6200/13 \;\; -2310/13)^T$$

Calculer l'erreur relative commise en utilisant la norme l_∞.

13. Pour la matrice de l'exercice précédent, calculer les quantités $||A||_1$, $||A||_2$, $||A||_\infty$, $||A^{-1}||_\infty$ et cond$A$.

14. Représenter graphiquement dans le plan les ensembles suivants:

 a) $\{\vec{x} \mid ||\vec{x}||_2 < 1\}$

 b) $\{\vec{x} \mid ||\vec{x}||_\infty < 1\}$

15. Est-ce que |dét A| pourrait être une norme matricielle?

16. Peut-on définir le conditionnement d'une matrice singulière?

17. Calculer le déterminant et le conditionnement de la matrice:

$$\begin{bmatrix} 1 & 2 \\ 1{,}01 & 2 \end{bmatrix}$$

18. Les matrices mal conditionnées ont souvent un déterminant voisin de 0. Est-ce que les matrices dont le déterminant est près de 0 sont forcément mal conditionnées? Donner un contre-exemple.

19. En utilisant un logiciel, calculer le conditionnement de la matrice de Hilbert de dimension 5, définie par:

$$a_{ij} = \frac{1}{i+j-1}$$

20. On considère le système linéaire:

$$\begin{bmatrix} 4 & 3 & -1 \\ 7 & -2 & 3 \\ 5 & -18 & 13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 9 \\ 3 \end{bmatrix}$$

En utilisant une calculatrice, on trouve comme solution le vecteur:

$$[1{,}586\,206 \quad -0{,}448\,2759 \quad -1{,}000\,000]^T$$

Après avoir multiplié la matrice par ce vecteur, on trouve:

$$[5{,}999\,9963 \;\; 8{,}999\,9938 \;\; 2{,}999\,9962]^T$$

ce qui semble indiquer que la solution obtenue est acceptable.

Le même calcul effectué sur ordinateur produit la solution:

$$[0{,}620\,6896 \quad 2{,}172\,4137 \quad 3{,}0000]^T$$

Après multiplication de la matrice A du système par ce vecteur, l'ordinateur affiche:

$$[5{,}999\,9995 \quad 8{,}999\,9998 \quad 3{,}000\,0014]^T$$

ce qui semble tout aussi acceptable. Que penser de ces résultats?

21. Considérer le système linéaire suivant:

$$\begin{pmatrix} 1 & 5 \\ 1{,}0001 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 6{,}0000 \\ 6{,}0005 \end{pmatrix}$$

dont la solution exacte est $\vec{x} = [5 \quad 0{,}2]^T$.

a) Calculer les résidus correspondant aux solutions approximatives $\vec{x}_1 = [5{,}1 \quad 0{,}3]^T$ et $\vec{x}_2 = [1 \quad 1]^T$. Calculer les quantités $||\vec{r}_1||_\infty$, $||\vec{r}_2||_\infty$, $||\vec{x} - \vec{x}_1||_\infty$ et $||\vec{x} - \vec{x}_2||_\infty$, puis comparer les résultats.

b) Trouver la solution exacte du système après le remplacement du membre de droite par $[6 \quad 6]^T$.

c) À la lueur des résultats obtenus en a) et en b), conclure sur le conditionnement de la matrice de ce système.

22. Résoudre le système non linéaire suivant à l'aide de la méthode de Newton en prenant $(0, 0)$ comme approximation initiale:

$$\begin{aligned} x_1^2 - 10x_1 + x_2^2 + 8 &= 0 \\ x_1x_2^2 + x_1 - 10x_2 + 8 &= 0 \end{aligned}$$

23. Résoudre le système non linéaire suivant à l'aide de la méthode de Newton en prenant $(0{,}1\,,\, 0{,}1\,,\, -0{,}1)$ comme approximation initiale:

$$\begin{aligned} 3x_1 - \cos(x_2x_3) - 1/2 &= 0 \\ x_1^2 - 81(x_2 + 0{,}1)^2 + \sin x_3 + 1{,}06 &= 0 \\ e^{-x_1x_2} + 20x_3 + (10\pi - 3)/3 &= 0 \end{aligned}$$

24. Tout comme dans le cas d'une équation d'une variable, la convergence de la méthode de Newton pour les systèmes non linéaires dépend de l'approximation initiale (x_1^0, x_2^0). Considérant le système:

$$\begin{aligned} x_2 + x_1^2 - x_1 - 1/2 &= 0 \\ x_2 + 5x_1x_2 - x_1^3 &= 0 \end{aligned}$$

expliquer pourquoi $(0\,,\,-0{,}2)$ est une mauvaise approximation initiale.

Chapitre 4

Systèmes dynamiques discrets

4.1 Introduction

Nous voyons dans ce chapitre comment de simples méthodes itératives, telles les méthodes de points fixes, peuvent mener à des systèmes au comportement complexe. On pourrait croire, à la suite du chapitre 2, que la discussion sur la convergence d'une méthode de points fixes s'arrête lorsqu'on a déterminé si le point fixe est attractif ou répulsif. Nous allons pousser cette discussion beaucoup plus loin et tâcher d'étudier un certain nombre de phénomènes intéressants rencontrés dans l'étude des systèmes dynamiques. Il ne s'agit pas de faire une analyse mathématique profonde de la théorie des systèmes dynamiques, mais bien de démontrer que des méthodes itératives simples peuvent résulter en des systèmes complexes.

4.2 Application quadratique

Nous reprenons ici une partie du travail de Feigenbaum (réf. [10]) sur l'application quadratique. Cette application remarquablement simple conduit à un comportement de nature universelle.

Considérons la méthode itérative:

$$\begin{cases} x_0 & \text{donné} \\ x_{n+1} = & \lambda x_n(1 - x_n) \end{cases} \tag{4.1}$$

qui est en fait une méthode de points fixes (voir l'équation 2.5) appliquée à la fonction:

$$g(x) = \lambda x(1 - x)$$

Le paramètre λ est appelé à varier, si bien que le comportement de l'algorithme 4.1 sera très différent suivant la valeur de λ.

Tout d'abord, il est facile de montrer que la fonction $g(x)$ est une application de l'intervalle $[0, 1]$ dans lui-même ($g(x) \in [0, 1]$ si $x \in [0, 1]$) seulement pour:

$$0 < \lambda < 4$$

En effet, le maximum de $g(x)$ est atteint en $x = 1/2$ et vaut $\lambda/4$. Nous nous limitons donc à ces valeurs de λ, qui sont de loin les plus intéressantes. En premier lieu, il convient de déterminer les points fixes de $g(x)$ et de vérifier s'ils sont attractifs ou répulsifs. Bien entendu, cela dépendra de λ. Les points fixes sont les solutions de:

$$x = g(x) = \lambda x(1 - x)$$

On constate immédiatement que 0 est une solution de cette équation et est donc un point fixe. Si on suppose que $x \neq 0$ et que l'on divise chaque côté de l'égalité par x, on obtient:

$$1 = \lambda(1 - x)$$

ce qui entraîne que:

$$x^* = \frac{\lambda - 1}{\lambda} \tag{4.2}$$

est un autre point fixe. En fait, 0 et x^* sont les deux seuls points fixes de $g(x)$. Voyons maintenant s'ils sont attractifs. Pour ce faire, il faut calculer la dérivée de $g(x)$, à savoir:

$$g'(x) = \lambda(1 - 2x) \tag{4.3}$$

On a donc d'une part:

$$g'(0) = \lambda$$

ce qui signifie que 0 sera un point fixe attractif si:

$$|g'(0)| < 1$$

c'est-à-dire si:

$$0 < \lambda < 1 = \lambda_1$$

puisqu'on ne considère pas les valeurs négatives de λ. D'autre part:

$$g'(x^*) = \lambda(1 - 2x^*) = \lambda(1 - 2(\frac{\lambda - 1}{\lambda})) = 2 - \lambda$$

Le point fixe x^* est donc attractif si:

$$|2 - \lambda| < 1$$

ou encore si:

$$\begin{array}{rcccl} -1 & < & 2-\lambda & < & 1 \\ -3 & < & -\lambda & < & -1 \\ 1 & < & \lambda & < & 3 = \lambda_2 \end{array}$$

On note λ_2 la borne supérieure de cet intervalle, soit $\lambda_2 = 3$. On en conclut que x^* est attractif pour ces valeurs de λ. On remarque de plus que, lorsque $\lambda = 1$, $x^* = 0$ et il n'y a alors qu'un seul point fixe. On montre que ce point fixe est attractif même si $g'(0) = 1$ en vertu de l'équation 4.3. La convergence est cependant très lente. La situation est illustrée aux figures 4.1 et 4.2, où la fonction $g(x)$ est tracée pour deux valeurs de λ, soit 0,5 et 1,5. On voit dans le premier cas ($\lambda = 0{,}5$) que la pente de $g(x)$ est inférieure à 1 en $x = 0$ et qu'il n'y a pas d'autre point fixe dans l'intervalle $[0, 1]$. Par contre, pour $\lambda = 1{,}5$, il y a bien deux points fixes $x = 0$ et $x = x^*$, dont seul x^* est attractif car $g'(x^*) < 1$.

Vérifions tout cela avec quelques exemples. Prenons d'abord $\lambda = 0{,}5$. On a alors:

$$x^* = -1$$

qui ne peut être attractif puisqu'il est l'extérieur de l'intervalle $[0, 1]$. À partir de $x_0 = 0{,}9$ par exemple, on trouve les itérations suivantes.

n	x_n
0	0,900 0000
1	0,045 0000
2	0,021 4875
3	0,010 5128
4	0,005 2011
5	0,002 5871
6	0,001 2902
7	0,000 6426
8	0,000 3219
9	0,000 1609
10	0,000 0804

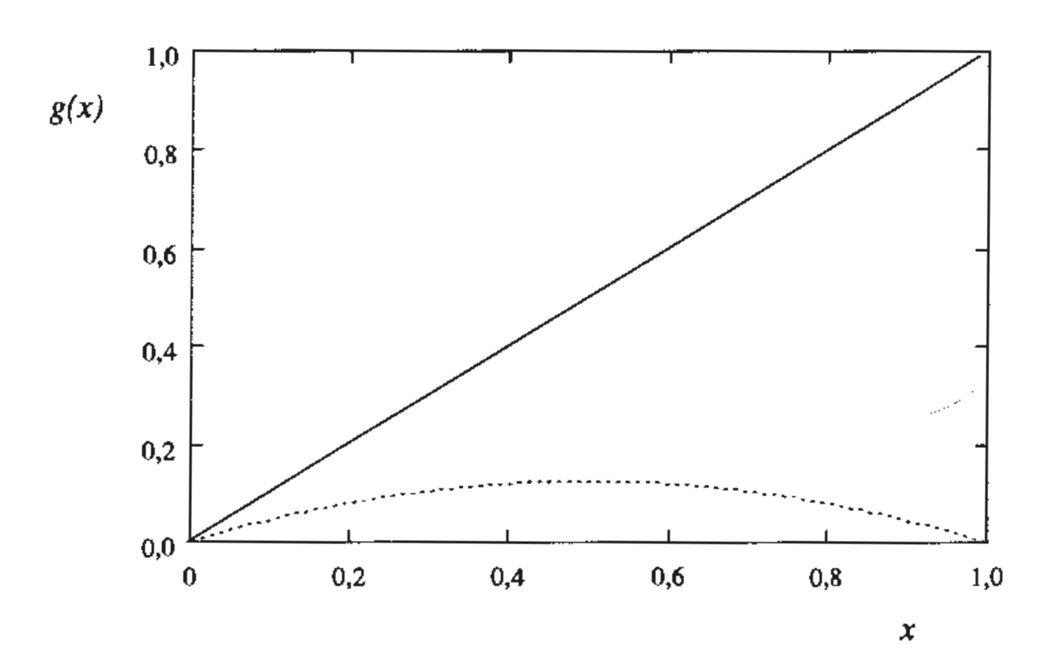


Figure 4.1: Application quadratique: $\lambda = 0{,}5$

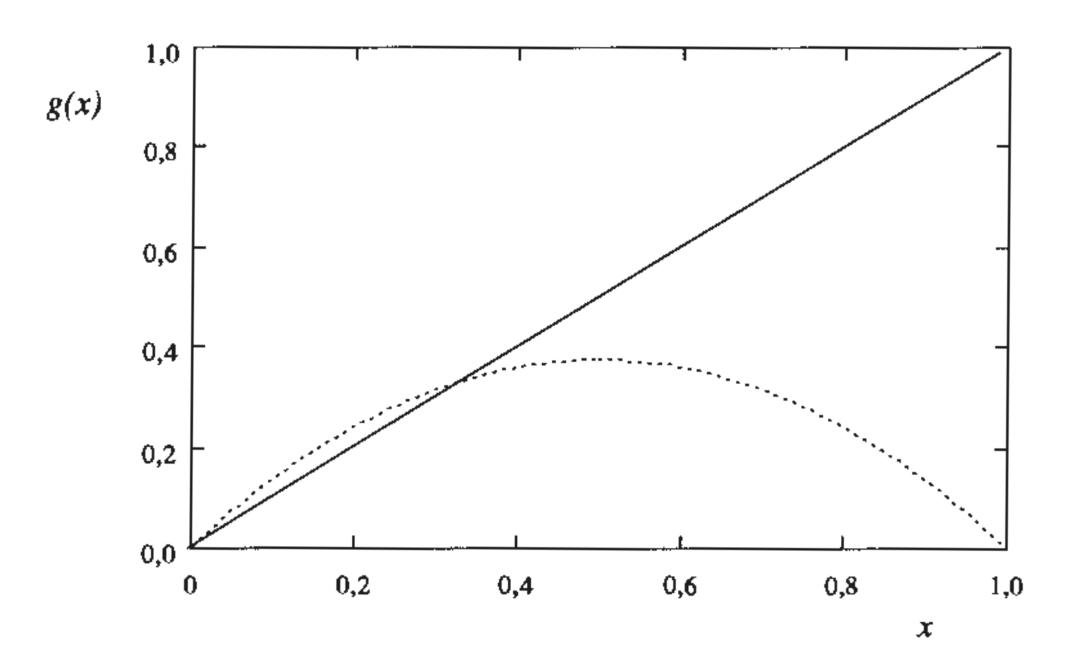


Figure 4.2: Application quadratique: $\lambda = 1{,}5$

Ces itérations convergent rapidement vers le point fixe 0. Si on prend maintenant $\lambda = 0{,}95$, toujours à partir de $x_0 = 0{,}9$, on obtient les itérations suivantes.

n	x_n
0	0,900 0000
1	0,085 5000
2	0,074 2802
3	0,065 3245
4	0,058 0044
5	0,051 9079
6	0,046 7527
7	0,042 3386
8	0,038 5187
9	0,035 1833
10	0,032 2482

Ces dernières convergent vers 0, mais beaucoup plus lentement. Cela tient au fait que le taux de convergence $g'(0) = \lambda$ vaut 0,5 dans le premier cas et 0,95 dans le second cas. La convergence est donc plus rapide pour $\lambda = 0{,}5$. Pour s'assurer de la convergence vers 0, il faudrait faire beaucoup plus que 10 itérations. Par exemple, pour $\lambda = 0{,}95$, on trouverait $x_{200} = 0{,}114\,1385 \times 10^{-5}$.

Passons maintenant à $\lambda = 1{,}5$, pour lequel $x^* = 1/3$. L'analyse a démontré que dans ce cas 0 est répulsif puisque $g'(0) = 1{,}5$, mais que x^* est attractif. À partir cette fois de $x_0 = 0{,}1$, on obtient les itérations suivantes.

n	x_n
0	0,100 0000
1	0,085 5000
2	0,074 2802
3	0,065 3245
4	0,058 0044
5	0,051 9079
6	0,046 7527
7	0,042 3386
8	0,038 5187
9	0,035 1833
10	0,032 2482
⋮	⋮
20	0,333 3313

Les itérations convergent donc vers $x^* = 1/3$, un résultat qui confirme l'analyse précédente. On obtiendrait des résultats similaires pour des valeurs de λ situées dans l'intervalle $]1, 3[$. Notons cependant que la valeur de x^* varie avec λ.

La question fondamentale est maintenant la suivante: que se passe-t-il si on prend des valeurs de λ supérieures à 3? On pourrait s'attendre à ce que les itérations de l'algorithme 4.1 divergent. Heureusement, ce n'est pas le cas, mais pour expliquer ce comportement il nous faut élargir la notion de convergence. Jusqu'à maintenant, nous n'avons parlé de convergence que vers un point (fixe). Or, il arrive qu'un algorithme converge vers autre chose qu'un point fixe. On parle alors d'un *attracteur* (voir par exemple Gulick, réf. [13]).

Définition 4.1
Un ensemble $A \subset R^n$ est dit un *attracteur* d'une application:

$$g : V \rightarrow R^n$$

où V est un sous-ensemble de R^n, si les conditions suivantes sont respectées:

1. Si $x \in A$, alors $g(x) \in A$.
2. Il existe un voisinage U de A tel que, si $x_0 \in U$, la suite $x_{n+1} = g(x_n)$ converge vers A.

Remarque 4.1
On ne considère pour l'instant que le cas $n = 1$ des applications de R dans R. Le cas général sera traité un peu plus loin. □

Remarque 4.2
La définition qui précède indique que, pour que A soit un attracteur, il faut que tout point x de l'ensemble A soit projeté sur un autre point de A par l'application $g(x)$. C'est bien sûr le cas d'un point fixe qui est envoyé sur lui-même. La deuxième condition traduit le fait que, si l'on part d'un point x_0 suffisamment près de A, les itérations de l'algorithme de points fixes s'approchent de plus en plus de l'ensemble A. □

Remarque 4.3
Un point fixe est donc un attracteur (voir le chapitre 2) s'il existe un intervalle I contenant ce point fixe pour lequel $g(x) \in I, \quad \forall x \in I$, et qui vérifie:

$$|g'(x)| < 1 \;\; \forall x \in I \qquad \square$$

Cette définition d'un attracteur est quelque peu imprécise, mais elle suffit aux besoins de l'exposé. Prenons par exemple $\lambda = 3{,}1$ et observons ce qui se passe. À partir de $x_0 = 0{,}5$, on obtient les itérations suivantes.

n	x_n
1	0,775 0000
2	0,540 5625
3	0,769 8995
4	0,549 1781
5	0,767 5026
6	0,553 1711
7	0,766 2357
8	0,555 2674
9	0,765 5310
10	0,556 4290
11	0,765 1288
12	0,557 0907
13	0,764 8960
14	0,557 4733
15	0,764 7601
16	0,557 6964
17	0,764 6804
18	0,557 8271
19	0,764 6336
20	0,557 9039
⋮	⋮
47	0,764 5665
48	0,558 0140
49	0,764 5665
50	0,558 0140

On remarque immédiatement un comportement surprenant. Les itérations oscillent entre les valeurs de 0,5580 et de 0,7645. Il n'y a donc pas convergence au sens habituel. En fait, les itérations paires convergent vers environ 0,558 014 et les itérations impaires, vers 0,764 566. Pour comprendre ce qui se passe, il suffit de constater que les itérations paires et impaires correspondent aux itérations de la fonction composée:

$$g_1(x) = g(g(x)) = \lambda g(x)(1 - g(x)) = \lambda(\lambda x(1 - x))(1 - \lambda x(1 - x))$$

c'est-à-dire

$$g_1(x) = \lambda^2 x(1 - x)(1 - \lambda x + \lambda x^2)$$

Pour déterminer les points fixes de la fonction $g_1(x)$, il suffit de résoudre:

$$x = g_1(x) = \lambda^2 x(1 - x)(1 - \lambda x + \lambda x^2)$$

Il est clair que tout point fixe de $g(x)$ est un point fixe de $g_1(x)$. Le point x^* donné par l'équation 4.2 ainsi que 0 sont donc des points fixes de $g_1(x)$, mais nous savons qu'ils sont répulsifs pour $\lambda > 3$. Il existe cependant d'autres points fixes de $g_1(x)$ qui ne sont pas des points fixes de $g(x)$. Après avoir divisé l'équation précédente par x de chaque côté, quelques manipulations algébriques nous amènent à résoudre l'équation:

$$\lambda^3 x^3 - 2\lambda^3 x^2 + \lambda^2(1 + \lambda)x + (1 - \lambda^2) = 0$$

dont les trois racines sont x^* et:

$$x_1, x_2 = \frac{1}{2} + \frac{1}{2\lambda} \pm \frac{1}{2\lambda}\sqrt{(\lambda - 3)(\lambda + 1)} \qquad (4.4)$$

On montre alors que:

$$x_1 = g(x_2) \quad \text{et} \quad x_2 = g(x_1)$$

c'est-à-dire que x_1 est envoyé sur x_2 par l'application $g(x)$ et vice versa. Dans le cas où $\lambda = 3{,}1$, on a en vertu de l'équation 4.4:

$$x_1 \simeq 0{,}558\,014 \quad \text{et} \quad x_2 \simeq 0{,}764\,566$$

ce qui correspond bien à ce que nous avons observé numériquement. L'ensemble $\{x_1, x_2\}$ est donc un attracteur au sens de notre définition. De plus, puisque les itérations oscillent entre deux valeurs, on parlera d'un *attracteur*

de période 2 ou d'un *2-cycle.* On peut dès lors s'interroger sur sa stabilité. En d'autres termes, pour quelles valeurs de λ ce 2-cycle est-il attractif?

Théorème 4.1

L'attracteur de période 2 (ou 2-cycle) donné par l'équation 4.4 est attractif pour:

$$3 < \lambda < \lambda_3 = 1 + \sqrt{6} \simeq 3{,}449\,489$$

Démonstration (voir Gulick, réf. [13]):
Il suffit de montrer que x_1 et x_2 sont des points fixes attractifs de la fonction $g_1(x) = g(g(x))$. La démonstration n'est faite que pour x_1, puisque l'autre cas est similaire. Par la règle de dérivation en chaîne, on a:

$$g_1'(x) = g'(g(x))g'(x)$$

de telle sorte que:

$$\begin{aligned} g_1'(x_1) &= g'(g(x_1))g'(x_1) = g'(x_2)g'(x_1) = \lambda^2(1 - 2x_1)(1 - 2x_2) \\ &= \lambda^2(1 - 2(x_1 + x_2) + 4x_1x_2) \end{aligned}$$

Un simple calcul à l'aide de l'équation 4.4 mène aux égalités:

$$x_1 + x_2 = 1 + \frac{1}{\lambda} \qquad x_1x_2 = \frac{1}{\lambda} + \frac{1}{\lambda^2}$$

de telle sorte que:

$$g_1'(x_1) = \lambda^2\left(1 - 2(1 + \frac{1}{\lambda}) + 4(\frac{1}{\lambda} + \frac{1}{\lambda^2})\right)$$

ou encore

$$g_1'(x_1) = -\lambda^2 + 2\lambda + 4$$

Il reste à obtenir les valeurs de λ pour lesquelles on aura:

$$|g_1'(x_1)| = |-\lambda^2 + 2\lambda + 4| < 1$$

C'est le cas si:

$$\begin{array}{rrcccl} & -1 & < & -\lambda^2+2\lambda+4 & < & 1 \\ \Leftrightarrow & -1 & < & -(\lambda-1)^2+5 & < & 1 \\ \Leftrightarrow & -6 & < & -(\lambda-1)^2 & < & -4 \\ \Leftrightarrow & 4 & < & (\lambda-1)^2 & < & 6 \\ \Leftrightarrow & 2 & < & (\lambda-1) & < & \sqrt{6} \\ \Leftrightarrow & 3 & < & \lambda & < & 1+\sqrt{6} \end{array} \qquad \square$$

On peut même démontrer que le cas $\lambda_3 = 1 + \sqrt{6}$ donne également lieu à un 2-cycle attractif (la démonstration est cependant délicate). Pour toute valeur de $\lambda \in]\lambda_2, \lambda_3] =]3, 1 + \sqrt{6}]$, on observe donc un 2-cycle attractif. La valeur des points x_1 et x_2 varie avec λ, mais le comportement général est le même. Le tableau suivant résume la convergence de l'application quadratique.

$0 < \lambda \leq \lambda_1 = 1$	0 est un point fixe attractif
$\lambda_1 < \lambda \leq \lambda_2 = 3$	x^* est un point fixe attractif
$\lambda_2 < \lambda \leq \lambda_3 = 1 + \sqrt{6}$	$\{x_1, x_2\}$ est un 2-cycle attractif

L'étude du cas où $\lambda > 1 + \sqrt{6}$ est relativement complexe, mais on peut en comprendre les grandes lignes. Le 2-cycle devient répulsif et est remplacé par un 4-cycle (2^2-cycle), qui est constitué des 4 points fixes $\{x_1, x_2, x_3, x_4\}$ de la fonction:

$$g_2(x) = g_1(g_1(x)) = g(g(g(g(x))))$$

vérifiant:

$$g(x_1) = x_2, \quad g(x_2) = x_3, \quad g(x_3) = x_4 \text{ et } g(x_4) = x_1$$

Ce 4-cycle est attractif pour $\lambda \in]\lambda_3, \lambda_4]$. Selon Gulick (réf. [13]), la valeur de λ_4 se situe autour de 3,539 58. À son tour, ce 4-cycle est remplacé par un 8-cycle (2^3-cycle) attractif dans l'intervalle $]\lambda_4, \lambda_5]$ et ainsi de suite. À chaque étape, la période de l'attracteur est doublée et on parle d'une *cascade de dédoublements de période.*

Plusieurs phénomènes intéressants se produisent alors:

- La distance entre les valeurs critiques λ_n où se produisent les dédoublements de période diminue et tend même vers 0. Cela signifie que l'intervalle où un 2^n-cycle sera attractif est d'autant plus étroit que n est grand. Il est par conséquent difficile d'observer expérimentalement un 2^n-cycle pour n grand.

- On peut montrer que:

$$\lim_{n \to \infty} \lambda_n = 3{,}615\,47...$$

- La distance entre les λ_n consécutifs est régie par ce qui semble une *loi universelle* de la forme:

$$\lim_{n \to \infty} \frac{\lambda_n - \lambda_{n-1}}{\lambda_{n+1} - \lambda_n} = d_\infty = 4{,}669\,202... \tag{4.5}$$

  Le nombre d_∞ est appelé *constante universelle de Feigenbaum.* La constante est universelle en ce sens que tout phénomène physique qui subit une cascade de dédoublements de période obéit à la loi 4.5. On a observé les cascades de dédoublements de période en laboratoire dans des problèmes de mécanique des fluides, de convection naturelle, de réactions chimiques, etc. Chaque fois, on a réussi à mettre au moins partiellement en évidence la relation 4.5.

- Pour les valeurs de λ supérieures à λ_∞, le comportement de l'application quadratique est très complexe. Sans en faire l'étude détaillée, on peut noter que les itérations successives se comportent parfois de façon périodique (de période souvent impaire) et parfois de façon chaotique. On dit d'un comportement qu'il est chaotique s'il n'est pas périodique et s'il donne l'apparence d'être aléatoire.

La situation complète est résumée à la figure 4.3, où l'on a placé en abscisse les valeurs de λ. On obtient cette figure en faisant 2000 itérations de l'algorithme 4.1 et en mettant en ordonnée les valeurs des 1000 dernières itérations seulement, et ce pour toutes les valeurs de λ dans l'intervalle $[2{,}8\,,\ 3{,}7]$ par incrément de 0,001. Dans le cas où les itérations convergent vers un point fixe, ces 1000 itérés se confondent en un seul point. Dans le cas où on a un 2-cycle attractif, les itérations se superposent sur deux points, etc. Les 1000 premières itérations permettent d'éliminer tout effet transitoire et de s'assurer que la convergence est bien établie. Un tel graphique est appelé *diagramme de bifurcation.*

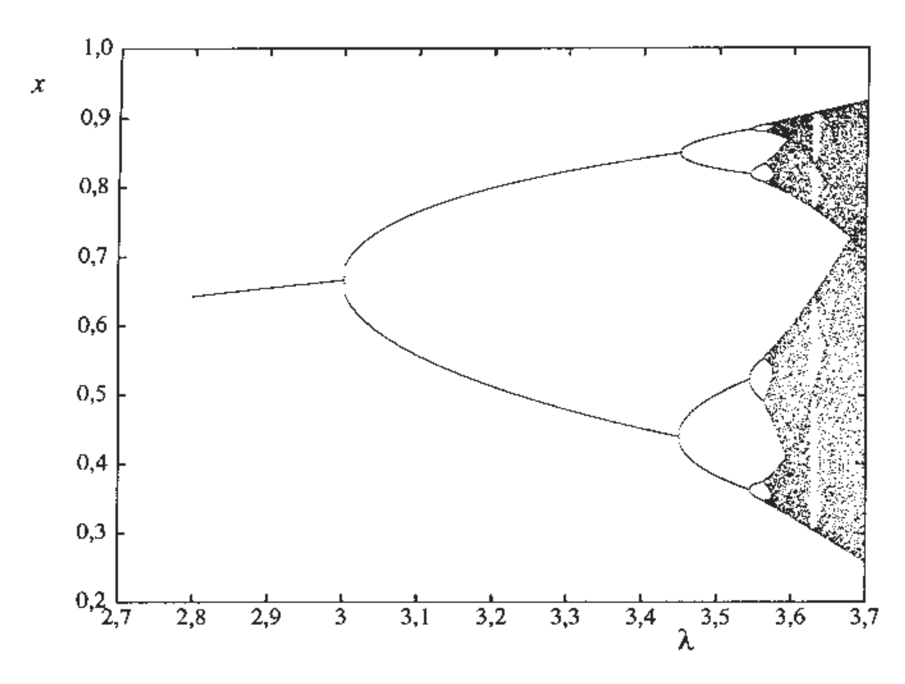


Figure 4.3: Application quadratique: diagramme de bifurcation

4.3 Méthodes de points fixes: cas complexe

On peut utiliser les méthodes de points fixes du chapitre 2 dans le cas d'équations non linéaires de la forme:

$$f(z) = 0 \tag{4.6}$$

où $z = x + iy$ est un nombre complexe et $f(z)$ est une fonction non linéaire d'une variable complexe. Dans ce cas précis, l'interprétation géométrique que nous avons donnée de cette méthode n'est plus valable, mais l'algorithme de base reste le même:

$$\begin{cases} z_0 & \text{donné} \\ z_{n+1} & = \quad g(z_n) \end{cases} \tag{4.7}$$

où $z_n = x_n + iy_n$.

Remarque 4.4
La dérivation d'une fonction d'une variable complexe suit les mêmes règles que celle des fonctions réelles. □

En particulier, la méthode de Newton s'écrit dans le cas complexe:

$$\begin{cases} z_0 & \text{donné} \\ z_{n+1} & = \quad z_n - \dfrac{f(z_n)}{f'(z_n)} \end{cases} \tag{4.8}$$

Exemple 4.1
On désire résoudre l'équation:

$$f(z) = z^2 + 1 = 0$$

qui ne possède évidemment pas de solution réelle, les deux solutions étant $z = \pm i$. L'algorithme devient dans ce cas:

$$z_{n+1} = z_n - \frac{(z_n^2 + 1)}{2z_n} = \frac{z_n}{2} - \frac{1}{2z_n}$$

À partir de $z_0 = 1 + i$, l'algorithme 4.8 donne les valeurs suivantes:

n	z_n
0	$1{,}0 + i$
1	$0{,}25 + i\,0{,}75$
2	$-0{,}75 + i\,0{,}975$
3	$0{,}001\,715 + i\,0{,}9973$
4	$0{,}928 \times 10^{-5} + i\,1{,}000\,002\,162$

On constate donc la convergence vers la solution $z = +i$.



Remarque 4.5
Dans certains langages informatiques, le calcul peut s'effectuer directement avec les nombres complexes. Il est également possible de séparer les parties réelles et imaginaires; on peut alors traiter seulement les nombres réels. On obtiendrait ainsi:

$$z_{n+1} = \frac{1}{2}\left(x_n - \frac{x_n}{(x_n^2 + y_n^2)}\right) + i\frac{1}{2}\left(y_n + \frac{y_n}{(x_n^2 + y_n^2)}\right)$$

où $z_n = x_n + iy_n$.
Si z_0 est choisi sur l'axe réel ($y_0 = 0$), l'algorithme diverge, car tous les z_n restent sur l'axe réel et ne peuvent jamais s'approcher de $z = \pm i$. □

La convergence des méthodes de points fixes appliquées aux variables complexes obéit à des règles similaires au cas réel. On a le résultat suivant.

Théorème 4.2
Soit $g(z)$, une fonction continue dans une région D du plan complexe et telle que $g(z) \in D$ pour tout z dans D. Si de plus $g'(z)$ existe et si:

$$|g'(z)| \leq k < 1$$

pour tout z dans D, alors tous les points z_0 de D appartiennent au bassin d'attraction de l'unique point fixe r de D. □

Remarque 4.6
Ce résultat est en parfaite concordance avec celui obtenu au chapitre 2, à la différence près que la norme de $g'(z)$ est la *norme complexe* (ou *module complexe*) définie par:

$$|z| = |x + iy| = \sqrt{x^2 + y^2} \qquad \square$$

Définition 4.2
La suite z_n définie par l'algorithme 4.7 est appelée l'*orbite du point* z_0.

En d'autres termes, l'orbite du point z_0 est la trajectoire du plan complexe que tracent les différents points z_n. Dans le dernier exemple, la trajectoire du point $z_0 = 1+i$ convergeait vers $z = i$. Cette définition nous entraîne vers un exemple très intéressant qui montre une fois de plus que des notions simples peuvent aboutir à des comportements complexes. Considérons le cas particulier de la fonction:

$$g_c(z) = z^2 + c \tag{4.9}$$

où $c = c_r + ic_i$ est un nombre complexe pour le moment quelconque. Pour une valeur de c donnée, l'orbite de $z_0 = 0$ aura typiquement deux comportements très différents l'un de l'autre. Pour certaines valeurs de c, les itérés

de l'algorithme 4.7 tendront vers l'infini. Par contre, pour d'autres valeurs de c, la suite z_n restera bornée. Cela nous permet de présenter l'*ensemble de Mandelbrot* (réf. [18]).

Définition 4.3
L'*ensemble de Mandelbrot* est défini comme étant l'ensemble des valeurs de c pour lesquelles l'orbite de $z_0 = 0$ pour l'algorithme de points fixes:

$$z_{n+1} = g_c(z_n) = z_n^2 + c$$

reste bornée lorsque $n \to \infty$.

Pour représenter l'ensemble de Mandelbrot, on utilise un algorithme très facile à programmer.

Algorithme 4.1: *Ensemble de Mandelbrot*

1. Étant donné N, le nombre maximal d'itérations
2. Étant donné M, une borne supérieure pour l'orbite
3. Étant donné $NPIXH$ et $NPIXV$, la résolution horizontale et verticale de l'écran utilisé
4. Considérer seulement les valeurs de $c = c_r + ic_i$ pour lesquelles:

$$c_r \in I_r = [c_r^1, c_r^2], \quad c_i \in I_i = [c_i^1, c_i^2]$$

5. Diviser les intervalles I_r et I_i en respectivement $NPIXH$ et $NPIXV$ sous-intervalles. Noter $c_{i,j}$, le point milieu du rectangle formé par le produit cartésien du i^e intervalle horizontal par le j^e intervalle vertical. Le point $c_{i,j}$ est ainsi associé au pixel situé à la i^e rangée et à la j^e colonne de l'écran.
6. Effectuer les opérations suivantes:
   (a) à partir de $z_0 = 0$
   (b) effectuer $z_1 = z_0^2 + c_{i,j}$
   (c) effectuer $z_0 = z_1$

(d) si $|z_0| > M$:
- orbite non bornée
- $c_{i,j}$ n'appartient pas à l'ensemble de Mandelbrot
- allumer le pixel associé à $c_{i,j}$
- passer à la valeur de $c_{i,j}$ suivante

(e) si $|z_0| < M$ et si le nombre maximal d'itérations N est atteint:
- orbite bornée
- $c_{i,j}$ appartient à l'ensemble de Mandelbrot
- pixel associé à $c_{i,j}$ reste non allumé (noir)
- passer à la valeur de $c_{i,j}$ suivante

(f) retour à l'étape (b) □

Voici quelques valeurs précises qui permettent d'obtenir de bons résultats.

- Le nombre maximal d'itérations N peut être fixé à 128.
- La borne supérieure M pour la norme de z_n peut être fixée à 4.
- On peut prendre les intervalles:

$$I_r = [-2,\ 1] \quad \text{et} \quad I_i = [-1{,}1,\ 1{,}1]$$

- On peut également introduire des variantes. On peut en effet colorer le pixel associé à une valeur de c qui n'appartient pas à l'ensemble de Mandelbrot en fonction du nombre d'itérations nécessaires pour que $|z_n| > M$.

On a obtenu la figure 4.4 en utilisant les valeurs précédentes. On remarque l'extrême irrégularité de la frontière de cet ensemble. Des agrandissements successifs de certaines parties permettraient de constater que le motif général se répète à l'infini.

4.4 Méthodes de points fixes en dimension n

Il est facile d'imaginer ce que peut être un point fixe dans le cas de plusieurs variables. Il s'agit simplement d'une solution de:

$$\vec{x} = \vec{g}(\vec{x}) \tag{4.10}$$

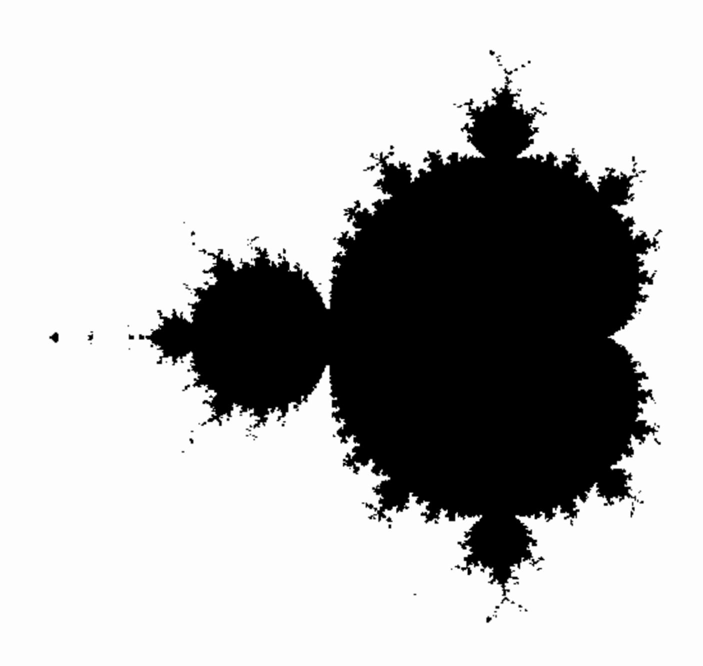


Figure 4.4: Ensemble de Mandelbrot

ou encore de

$$\begin{array}{rcl} x_1 & = & g_1(x_1, x_2, x_3, \cdots, x_n) \\ x_2 & = & g_2(x_1, x_2, x_3, \cdots, x_n) \\ x_3 & = & g_3(x_1, x_2, x_3, \cdots, x_n) \\ \vdots & & \vdots \\ x_n & = & g_n(x_1, x_2, x_3, \cdots, x_n) \end{array} \tag{4.11}$$

Définition 4.4
Tout vecteur $\vec{r}$ solution du système 4.10 ou 4.11 est appelé *point fixe* de l'application $\vec{g}(\vec{x})$.

Remarque 4.7
Nous ne faisons pas la distinction entre le point $(x_1, x_2, x_3, \cdots, x_n)$ et le vecteur $\vec{x} = [x_1 \;\; x_2 \;\; x_3 \; \cdots \; x_n]^T$. Les deux notations sont utilisées indifféremment dans cette section. □

L'algorithme de base des méthodes de points fixes en dimension n reste le même:

$$\begin{cases} \vec{x}^0 & \text{donné,} \\ \vec{x}^{i+1} = & \vec{g}(\vec{x}^i) \end{cases} \tag{4.12}$$

où $\vec{x}^i = [x_1^i \; x_2^i \; x_3^i \; \cdots \; x_n^i]^T$. Bien que présentant des similitudes avec le cas unidimensionnel, l'analyse de convergence des méthodes de points fixes en dimension n est cependant beaucoup plus délicate. C'est pourquoi nous préférons procéder par analogie avec le cas unidimensionnel. Nous avons vu que la convergence vers une racine r de l'algorithme des points fixes est assujettie à la condition:

$$|g'(r)| < 1$$

Il faut donc trouver l'expression analogue de cette condition en dimension n. Pour y parvenir, on doit revoir les notions de valeurs et de vecteurs propres qui seront à la base de la convergence de l'algorithme 4.12.

Définition 4.5
Si A est une matrice de dimension n, on définit le *polynôme caractéristique* de A par:

$$p(\lambda) = \text{dét } (A - \lambda I) \tag{4.13}$$

où I est la matrice identité.

Le polynôme $p(\lambda)$ est de degré n et possède donc n racines réelles ou complexes conjuguées.

Définition 4.6
Les racines (ou zéros) du polynôme caractéristique sont appelées *valeurs propres* de la matrice A.

Si λ est une valeur propre, la matrice $A - \lambda I$ est singulière (puisque son déterminant est nul) et le système:

$$(A - \lambda I)\vec{x} = 0$$

ou encore

$$A\vec{x} = \lambda I \vec{x} = \lambda \vec{x} \tag{4.14}$$

possède des solutions non nulles. En effet, le système 4.14 possède toujours la solution $\vec{x} = \vec{0}$; si λ est une valeur propre, il existe également d'autres solutions.

Définition 4.7
Une solution non nulle du système 4.14 est appelée *vecteur propre* de A associé à la valeur propre λ.

Exemple 4.2
Soit la matrice:

$$\begin{bmatrix} 1 & 2 \\ -2 & 2 \end{bmatrix}$$

Le polynôme caractéristique est alors:

$$\text{dét } (A - \lambda I) = \begin{vmatrix} 1-\lambda & 2 \\ -2 & 2-\lambda \end{vmatrix} = (1-\lambda)(2-\lambda) + 4 = \lambda^2 - 3\lambda + 6$$

Les racines de ce polynôme sont:

$$\lambda_{1,2} = \frac{3 \pm i\sqrt{15}}{2}$$

qui sont donc les 2 valeurs propres de la matrice A.

● ● ● ● ●

Remarque 4.8
Ce sont uniquement les valeurs propres qui sont importantes dans ce qui suit. Pour cette raison, nous ne nous attardons pas au calcul des vecteurs propres. □

Revenons aux méthodes de points fixes par le biais d'un cas particulier de l'équation 4.10 de la forme:

$$\vec{x} = A\vec{x} \qquad \text{(c'est-à-dire } \vec{g}(\vec{x}) = A\vec{x}\text{)}$$

où A est une matrice de dimension n sur n quelconque. On note que $\vec{0}$ est toujours un point fixe et que l'algorithme de points fixes 4.12 prend la forme:

$$\vec{x}^{i+1} = A\vec{x}^{i} = A(A\vec{x}^{i-1}) = \cdots = A^{i+1}\vec{x}^{0} \tag{4.15}$$

Il y aura convergence vers $\vec{0}$ de la méthode de points fixes si la suite $A^i\vec{x}^0$ tend vers $\vec{0}$ lorsque i tend vers l'infini. Il reste donc à établir sous quelles conditions cela se produit. Pour ce faire, il nous faut introduire deux autres définitions.

Définition 4.8
La *rayon spectral* d'une matrice A est défini par:

$$\rho(A) = \max_{1 \leq i \leq n} |\lambda_i| \tag{4.16}$$

où $|\lambda_i|$ est le module complexe de la valeur propre λ_i de A.

Remarque 4.9
Le calcul du rayon spectral requiert le calcul de la plus grande valeur propre en module. Certaines techniques permettent de déterminer le rayon spectral, mais ce n'est pas un problème facile surtout dans le cas des matrices de grande taille. □

Définition 4.9
Une matrice A est dite *convergente* si:

$$\lim_{n\to\infty} A^n = 0 \tag{4.17}$$

Selon cette dernière définition, il est clair que l'algorithme de points fixes 4.15 convergera vers le vecteur $\vec{0}$ si la matrice A est convergente. Le théorème suivant fournit des conditions équivalentes et permet de déterminer si une matrice est convergente.

Théorème 4.3
Les conditions suivantes sont équivalentes.

1. La matrice A est convergente.
2. Pour toute norme matricielle:
$$\lim_{n \to \infty} ||A^n|| = 0 \tag{4.18}$$
3. Pour tout vecteur $\vec{x}$:
$$\lim_{n \to \infty} A^n \vec{x} = 0 \tag{4.19}$$
4. Le rayon spectral de A est strictement inférieur à 1 ($\rho(A) < 1$). □

Le lecteur intéressé trouvera la démonstration complète de ce résultat dans Varga (réf. [26]). Il est assez facile de se convaincre de l'équivalence des 3 premiers énoncés. La partie la plus exigeante de la démonstration consiste à montrer que la matrice est convergente lorsque le rayon spectral est iuférieur à 1.

Remarque 4.10
Le théorème précédent permet d'affirmer que $\vec{0}$ est un point fixe attractif de l'algorithme 4.15 si et seulement si le rayon spectral de la matrice A est inférieur à 1. Dans ce cas particulier, l'algorithme convergera quel que soit le vecteur initial $\vec{x}^0$ choisi, en vertu des relations 4.15 et 4.19.□

Exemple 4.3
Soit la matrice:
$$A = \begin{bmatrix} 1/2 & 0 \\ 1/3 & 1/4 \end{bmatrix}$$
dont les valeurs propres sont tout simplement $\lambda_1 = 1/2$ et $\lambda_2 = 1/4$. Un simple calcul permet de s'assurer que:
$$A^2 = \begin{bmatrix} 1/4 & 0 \\ 1/4 & 1/16 \end{bmatrix} \quad \cdots \quad A^{10} = \begin{bmatrix} 0{,}976\,56 \times 10^{-3} & 0{,}0 \\ 0{,}130\,08 \times 10^{-2} & 0{,}953\,67 \times 10^{-6} \end{bmatrix}$$

et que:

$$A^{50} = \begin{bmatrix} 0{,}888\,18 \times 10^{-15} & 0{,}0 \\ 0{,}118\,42 \times 10^{-14} & 0{,}788\,886 \times 10^{-30} \end{bmatrix}$$

On constate que chaque coefficient de la matrice A^n tend vers 0. □

• • • • •

Nous avons maintenant en main les outils nécessaires pour aborder le cas général de l'algorithme 4.12. L'exercice consiste à déterminer sous quelles conditions un point fixe noté $\vec{r}$ d'un système de dimension n est attractif.

On sait qu'en dimension 1 la convergence vers le point fixe r est liée à $g'(r)$, dout le rôle en dimension n est joué par la matrice jacobieune:

$$J(\vec{r}) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1}(\vec{r}) & \frac{\partial g_1}{\partial x_2}(\vec{r}) & \cdots & \frac{\partial g_1}{\partial x_n}(\vec{r}) \\ \frac{\partial g_2}{\partial x_1}(\vec{r}) & \frac{\partial g_2}{\partial x_2}(\vec{r}) & \cdots & \frac{\partial g_2}{\partial x_n}(\vec{r}) \\ \vdots & \vdots & \ddots & \\ \frac{\partial g_n}{\partial x_1}(\vec{r}) & \frac{\partial g_n}{\partial x_2}(\vec{r}) & \cdots & \frac{\partial g_n}{\partial x_n}(\vec{r}) \end{bmatrix}$$

L'équivalent multidimensionnel de la condition:

$$|g'(r)| < 1$$

est tout simplement:

$$\rho(J(\vec{r})) < 1$$

Le théorème suivant, démontré dans Burden et Faires (réf. [2]), résume la situatiou.

Théorème 4.4

Soit $\vec{r}$, un point fixe de l'application:

$$\vec{x} = \vec{g}(\vec{x})$$

$\vec{r}$ est attractif si le rayon spectral $\rho(J(\vec{r}))$ de la matrice jacobienne de $\vec{g}(\vec{x})$ est inférieur à 1; il est répulsif si $\rho(J(\vec{r})) > 1$. Le cas où $\rho(J(\vec{r})) = 1$ est indéterminé. □

Remarque 4.11
Ce théorème est une généralisation du cas unidimensionnel, auquel cas la matrice jacobienne se réduit à la matrice 1×1, dont le seul coefficient est $g'(r)$. L'unique valeur propre est bien sûr $g'(r)$, de telle sorte que le rayon spectral est $|g'(r)|$. Ce dernier doit être inférieur à 1 pour que le point fixe r soit attractif. □

Remarque 4.12
Dans le cas général comme dans le cas unidimensionnel, le fait que le point fixe $\vec{r}$ soit attractif ne garantit pas la convergence de l'algorithme 4.12. On doit toujours s'assurer que le vecteur initial $\vec{x}^0$ appartient au bassin d'attraction dn point fixe. La tâche est d'autant plus difficile que le système est de grande dimension. □

Exemple 4.4
Il est facile de démontrer que l'application:

$$\begin{aligned} x_1 &= \sqrt{2-(x_2)^2} \\ x_2 &= \sqrt{x_1} \end{aligned}$$

ne possède que le seul point fixe $[1 \;\; 1]^T$ (voir les exercices de fin de chapitre). La matrice jacobienne est dans ce cas:

$$J(x_1, x_2) = \begin{bmatrix} 0 & \frac{-x_2}{\sqrt{2-(x_2)^2}} \\ \frac{1}{2\sqrt{x_1}} & 0 \end{bmatrix}$$

qui, évaluée en $\vec{r} = [1 \;\; 1]^T$, donne:

$$J(1,1) = \begin{bmatrix} 0 & -1 \\ 1/2 & 0 \end{bmatrix}$$

Le polynôme caractéristique est alors:

$$\lambda^2 + 1/2 = 0$$

et les valeurs propres sont $\pm i\sqrt{1/2}$. Le point fixe $[1\ 1]^T$ est attractif puisque le rayon spectral est $\sqrt{1/2}$ et est inférieur à 1.

Si on retient le vecteur $\vec{x}^0 = [0 \quad 0]^T$ comme solution initiale, on trouve les valeurs suivantes de l'algorithme 4.12.

n	x_1^n	x_2^n
1	$0{,}141\,42 \times 10^1$	$0{,}000\,00 \times 10^0$
2	$0{,}141\,42 \times 10^1$	$0{,}118\,92 \times 10^1$
3	$0{,}765\,37 \times 10^0$	$0{,}118\,92 \times 10^1$
4	$0{,}765\,37 \times 10^0$	$0{,}874\,85 \times 10^0$
5	$0{,}111\,11 \times 10^1$	$0{,}874\,85 \times 10^0$
6	$0{,}111\,11 \times 10^1$	$0{,}105\,41 \times 10^1$
7	$0{,}942\,79 \times 10^0$	$0{,}105\,41 \times 10^1$
8	$0{,}942\,79 \times 10^0$	$0{,}970\,98 \times 10^0$
9	$0{,}102\,82 \times 10^1$	$0{,}970\,98 \times 10^0$
10	$0{,}102\,82 \times 10^1$	$0{,}101\,40 \times 10^1$
11	$0{,}985\,80 \times 10^0$	$0{,}101\,40 \times 10^1$
12	$0{,}985\,80 \times 10^0$	$0{,}992\,87 \times 10^0$
13	$0{,}100\,71 \times 10^1$	$0{,}992\,87 \times 10^0$
14	$0{,}100\,71 \times 10^1$	$0{,}100\,35 \times 10^1$
15	$0{,}996\,46 \times 10^0$	$0{,}100\,35 \times 10^1$
16	$0{,}996\,46 \times 10^0$	$0{,}998\,23 \times 10^0$
17	$0{,}100\,18 \times 10^1$	$0{,}998\,23 \times 10^0$
18	$0{,}100\,18 \times 10^1$	$0{,}100\,09 \times 10^1$
19	$0{,}999\,11 \times 10^0$	$0{,}100\,09 \times 10^1$
20	$0{,}999\,11 \times 10^0$	$0{,}999\,56 \times 10^0$
21	$0{,}100\,04 \times 10^1$	$0{,}999\,56 \times 10^0$
22	$0{,}100\,04 \times 10^1$	$0{,}100\,02 \times 10^1$
23	$0{,}999\,78 \times 10^0$	$0{,}100\,02 \times 10^1$
24	$0{,}999\,78 \times 10^0$	$0{,}999\,89 \times 10^0$
25	$0{,}100\,01 \times 10^1$	$0{,}999\,89 \times 10^0$

On constate une convergence relativement lente vers le point fixe $[1 \quad 1]^T$.

• • • • •

4.4.1 Attracteur d'Hénon

Considérons un cas bien particulier d'application non linéaire:

$$\begin{array}{rcl} g_1(x_1, x_2) & = & 1 - a(x_1)^2 + x_2 \\ g_2(x_1, x_2) & = & bx_1 \end{array} \tag{4.20}$$

où les paramètres a et b sont précisés plus loin. L'algorithme de points fixes en dimension 2 devient dans ce cas:

$$\left\{ \begin{array}{rcl} (x_1^0, x_2^0) & & \text{donné} \\ & & \\ x_1^{n+1} & = & 1 - a(x_1^n)^2 + x_2^n \\ x_2^{n+1} & = & bx_1^n \end{array} \right. \tag{4.21}$$

Déterminons en premier lieu les points fixes de cette application. Il suffit de résoudre:

$$\begin{array}{rcl} x_1 & = & 1 - ax_1^2 + x_2 \\ x_2 & = & bx_1 \end{array}$$

En substituant bx_1 à x_2 dans la première équation, on conclut que cette application possède deux points fixes:

$$\vec{r}_1 = \begin{pmatrix} \frac{1}{2a}(b - 1 + \sqrt{(1-b)^2 + 4a}) \\ \\ \frac{b}{2a}(b - 1 + \sqrt{(1-b)^2 + 4a}) \end{pmatrix}$$

$$\vec{r}_2 = \begin{pmatrix} \frac{1}{2a}(b - 1 - \sqrt{(1-b)^2 + 4a}) \\ \\ \frac{b}{2a}(b - 1 - \sqrt{(1-b)^2 + 4a}) \end{pmatrix}$$

La matrice jacobienne de $\vec{g}(\vec{x})$ est:

$$J(x_1, x_2) = \begin{bmatrix} -2ax_1 & 1 \\ \\ b & 0 \end{bmatrix}$$

dont les valeurs propres sont:

$$\lambda = -ax_1 \pm \sqrt{a^2x_1^2 + b}$$

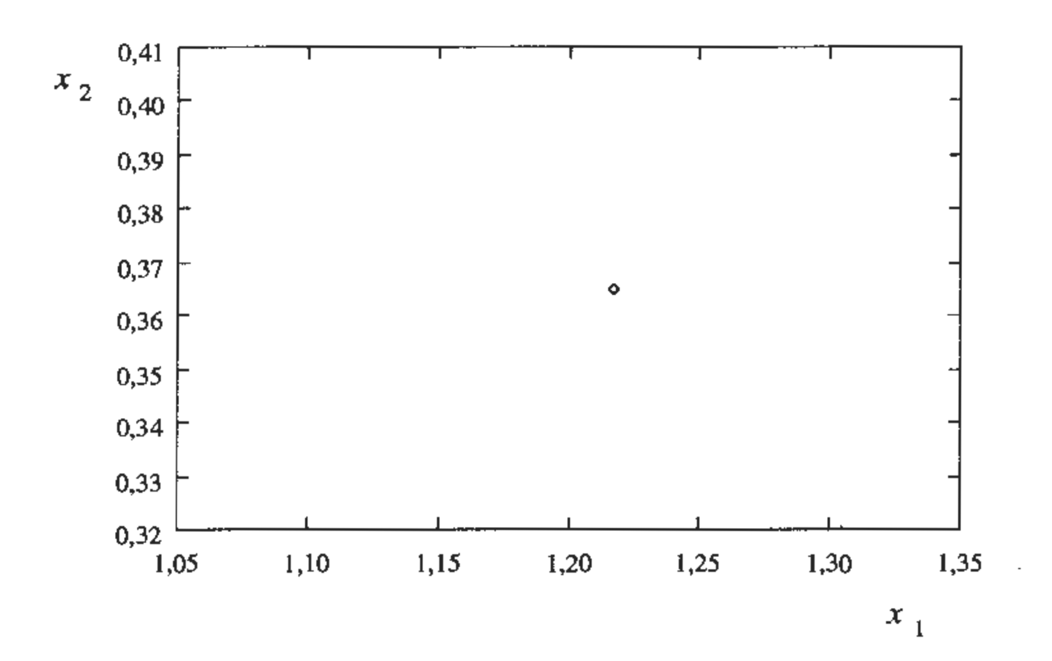


Figure 4.5: Application d'Hénon: $a = 0{,}1$

Il est alors possible de démontrer (voir Gulick, réf. [13]) que $\vec{r}_1$ est attractif pour des valeurs de a et b satisfaisant la condition:

$$-\frac{1}{4}(1-b)^2 < a < \frac{3}{4}(1-b)^2 \qquad (a \neq 0)$$

Le point fixe $\vec{r}_2$ est répulsif.

Ce qui suit ressemble beaucoup à ce que nous avons vu au sujet de l'application quadratique. Si on fixe $b = 0{,}3$ et si on fait varier le paramètre a, $\vec{r}_1$ est attractif pour les valeurs de a dans l'intervalle $]-0{,}1225\,,\, 0{,}3675[$. Par exemple, pour $a = 0{,}1$, le point fixe prend la valeur approximative $\vec{r}_1 = [1{,}211\,699 \;\; 0{,}365\,097]^T$. À partir de $\vec{x}^0 = [1 \;\; 1]^T$, l'algorithme de points fixes 4.12 a donné les résultats de la figure 4.5. On note la convergence vers le point fixe $\vec{r}_1 = [1{,}211\,699 \;\; 0{,}365\,097]^T$. Cette figure est le produit des 9000 dernières itérations effectuées sur un total de 10 000 itérations. Le seul point visible est en fait constitué de 9000 points superposés.

Si on augmente la valeur de a à 0,5 (voir la figure 4.6), on remarque la convergence vers un attracteur périodique de période 2 (2-cycle) constitué de deux points. Si on augmente progressivement la valeur de a, le comportement observé pour l'application quadratique se répète et on passe

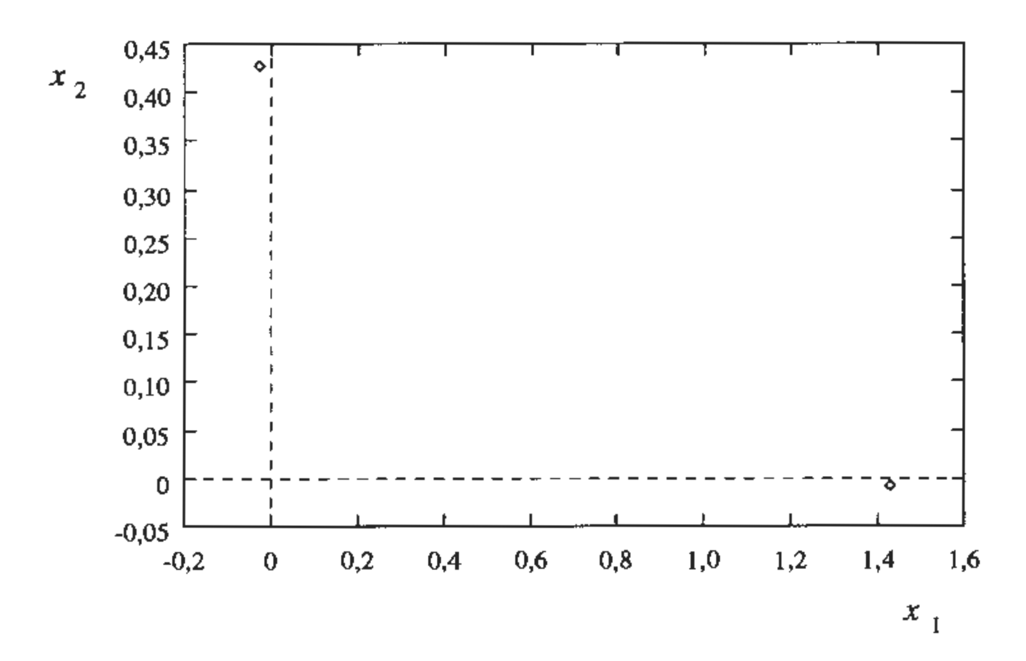


Figure 4.6: Application d'Hénon: $a = 0{,}5$

successivement à des attracteurs de période 4, 8, 16, 32, etc. Selon Derrida, Gervois et Pomeau (réf. [8]), les valeurs précises de a où se produisent les doublements de période sont les suivantes.

a	Période
$-0{,}1225$	1
0,3675	2
0,9125	4
1,0260	8
1,0510	16
1,0565	32
⋮	⋮

La figure 4.7 illustre l'attracteur de période 16. Il est intéressant de constater que:

$$\frac{a_5 - a_4}{a_6 - a_5} = \frac{1{,}0510 - 1{,}0260}{1{,}0565 - 1{,}0510} = 4{,}5454$$

ce qui est une bonne approximation de la constante universelle de Feigenbaum (voir l'équation 4.5). Si on se donnait la peine de localiser précisément

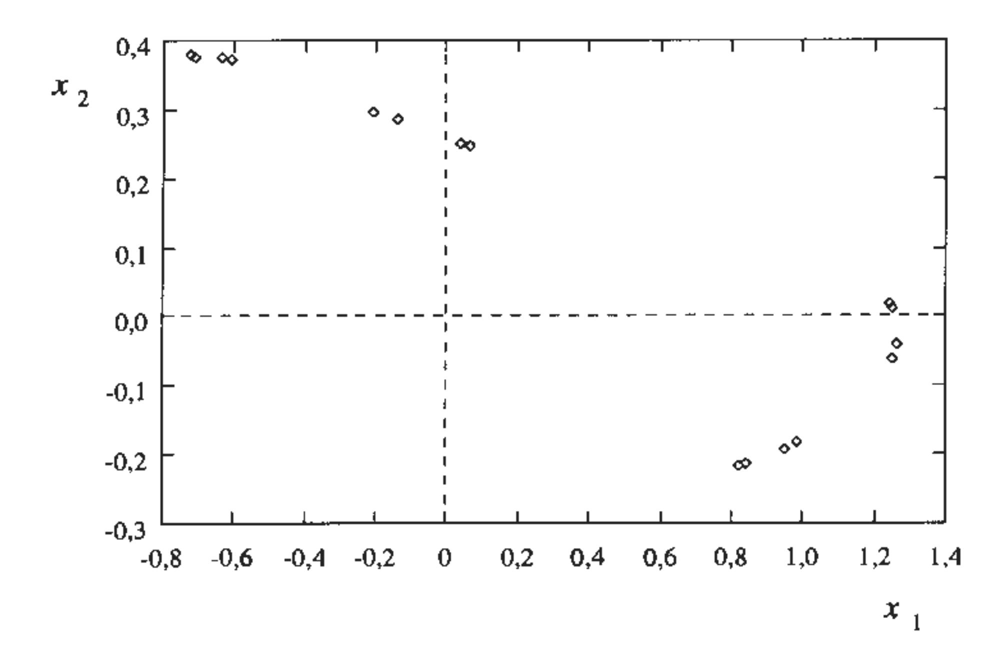


Figure 4.7: Application d'Hénon: $a = 1{,}054$

les valeurs de a où apparaissent les attracteurs de période 64, 128, 256, etc., on verrait ressortir la loi universelle 4.5. La cascade de dédoublements de période s'arrête autour de $a = 1{,}058\,0459$.

Si on continue à augmenter la valeur de a, on observe un comportement général de plus en plus complexe. Par exemple, pour $a = 1{,}4$, les itérations convergent vers l'attracteur présenté à la figure 4.8, et ce quel que soit le point de départ de l'algorithme. Ce qui est plus intéressant encore, c'est que les itérés (x_1^n, x_2^n) de l'algorithme de points fixes 4.21 parcourent cet attracteur de façon apparemment aléatoire, sans jamais s'en éloigner. On parle alors d'un *attracteur étrange.*

La complexité géométrique d'un tel attracteur est révélée par les agrandissements successifs de certaines parties de l'attracteur d'Hénon. Ce qui semblait être un trait gras sur la figure 4.8 devient, si on regarde de plus près, une série de 6 courbes sur la figure 4.9. Enfin, si on agrandit une section de ce qui semble être 3 courbes à la figure 4.9, on voit poindre 6 autres courbes à la figure 4.10. Ce phénomène peut théoriquement se reproduire à l'infini. Mais, pour le vérifier, il faudrait calculer beaucoup plus que les 50 000 points illustrés dans les figures mentionnées.

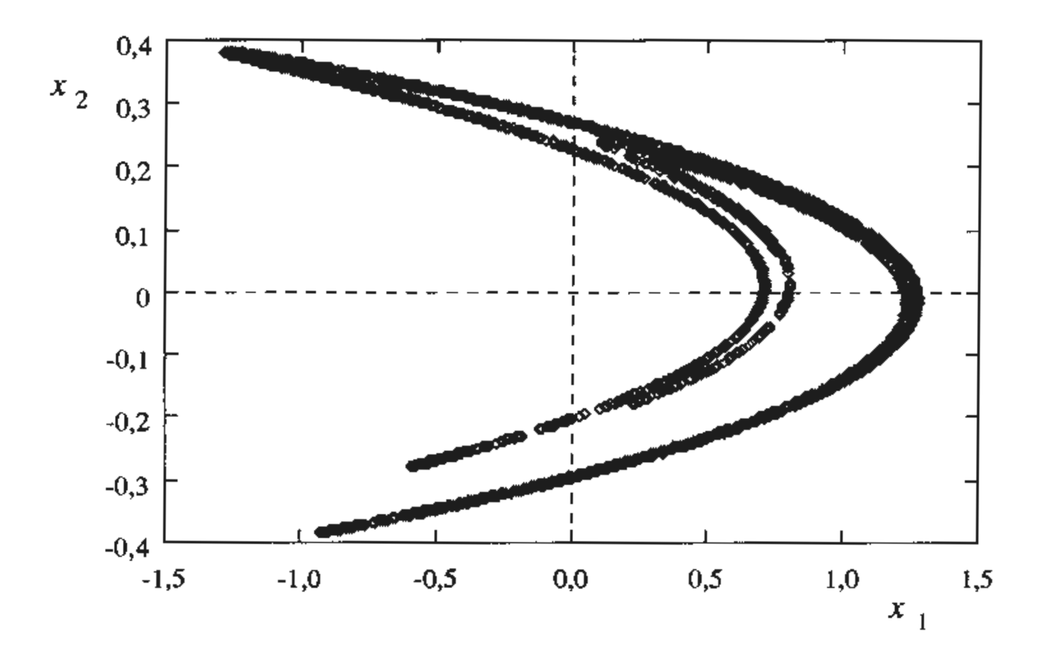


Figure 4.8: Attracteur d'Hénon: $a = 1{,}4$

L'attracteur d'Hénon est en fait un objet de *dimension fractionnaire* ou *fractale*. En effet, la dimension (terme qu'il faudrait préciser) de cet ensemble est supérieure à 1, mais strictement inférieure à 2. Les figures précédentes montrent que l'attracteur d'Hénon est plus qu'une courbe (de dimension 1), sans toutefois être une surface (qui serait de dimension 2). On retrouve une situation analogue à la frontière de l'ensemble de Mandelbrot.

4.5 Méthodes itératives pour les systèmes linéaires

La résolution numérique des grands systèmes linéaires peut parfois nécessiter l'emploi de méthodes autres que la décomposition LU. La raison principale est que la décomposition LU requiert la mise en mémoire d'une matrice de très grande taille, avec peu de possibilités de comprimer toute cette information. Les méthodes itératives, en revanche, permettent de ne placer en mémoire que les coefficients non nuls d'une matrice. Cela est particulièrement important avec les *matrices creuses*, dont une grande partie des coefficients sont nuls. La décomposition LU ne permet pas cette possibilité puisque le processus même de décomposition tend à remplir la matrice.

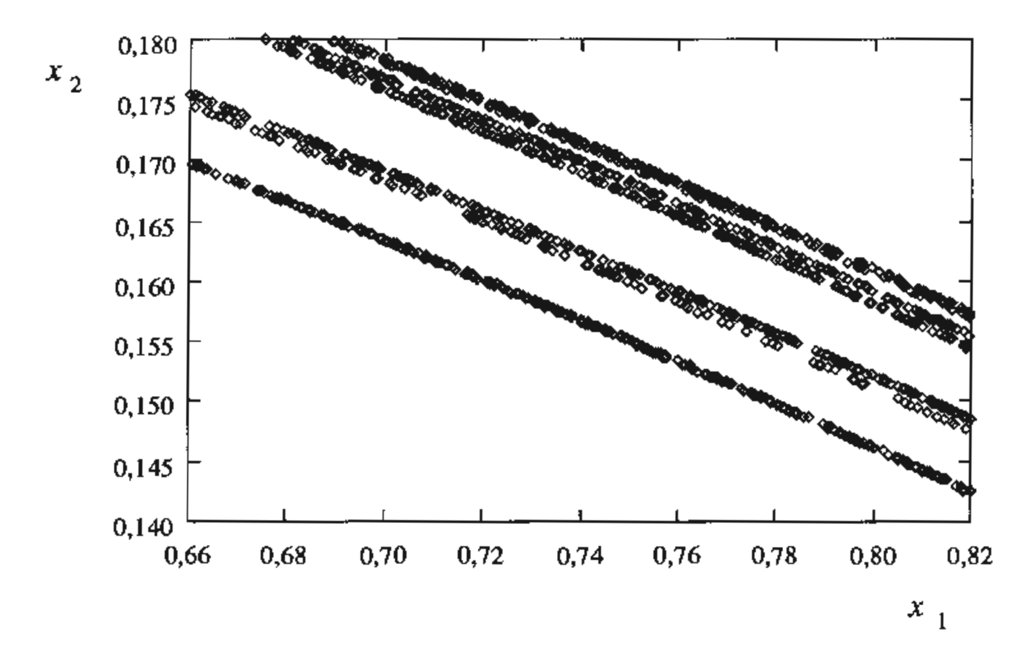


Figure 4.9: Attracteur d'Hénon: $a = 1,4$

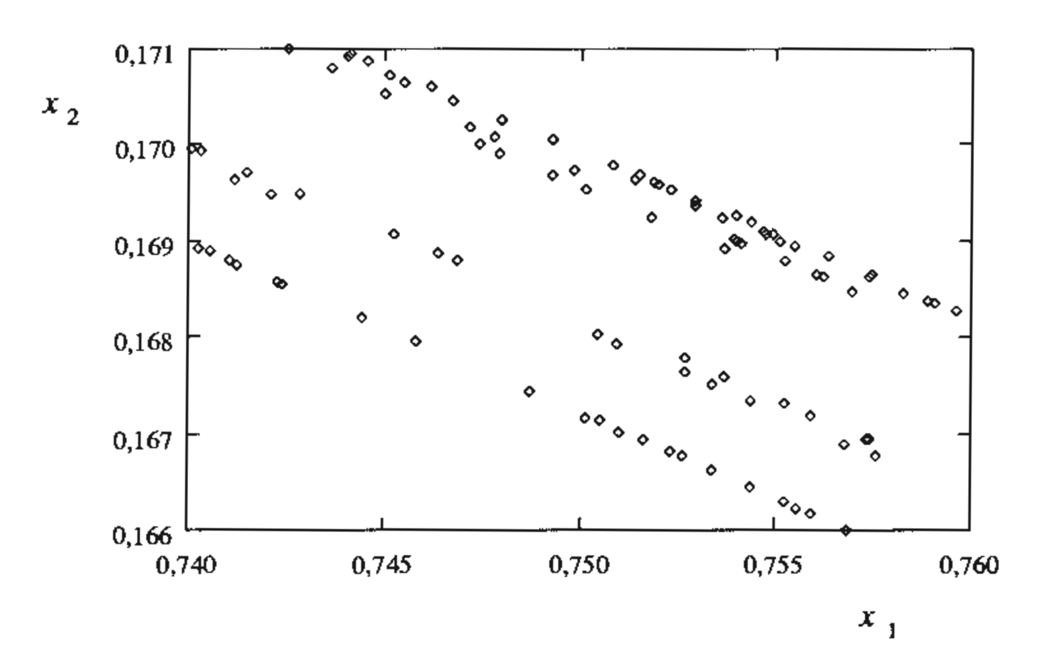


Figure 4.10: Attracteur d'Hénon: $a = 1,4$

En effet, la plupart des coefficients nuls d'une matrice creuse deviennent non nuls au terme de la décomposition.

Les méthodes itératives possèdent donc des avantages suffisamment importants pour justifier une recherche active dans ce domaine. Cependant, contrairement à la décomposition LU, le succès n'est pas assuré quelle que soit la matrice A pour laquelle on souhaite résoudre un système linéaire de la forme:

$$A\vec{x} = \vec{b}$$

La convergence des méthodes itératives n'est réalisée que dans certaines conditions que nous préciserons. Une grande prudence est donc de mise. De plus, les méthodes itératives, lorsqu'elles convergent, ne deviennent vraiment avantageuses que pour les systèmes linéaires de très grande taille.

4.5.1 Méthode de Jacobi

Considérons le système linéaire:

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n & = & b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots + a_{3n}x_n & = & b_3 \\ \vdots & = & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n & = & b_n \end{array} \qquad (4.22)$$

On suppose pour l'instant que tous les éléments de la diagonale sont non nuls ($a_{ii} \neq 0, \forall i$). À partir d'une approximation initiale de la solution que nous noterons $[x_1^0 \ x_2^0 \ \cdots \ x_n^0]^T$ (comme dans toute méthode itérative),

on construit l'algorithme:

$$\begin{aligned}
x_1^{k+1} &= \frac{1}{a_{11}}\left(b_1 - \sum_{j=2}^{n} a_{1j}x_j^k\right) \\
x_2^{k+1} &= \frac{1}{a_{22}}\left(b_2 - \sum_{j=1,j\neq 2}^{n} a_{2j}x_j^k\right) \\
x_3^{k+1} &= \frac{1}{a_{33}}\left(b_3 - \sum_{j=1,j\neq 3}^{n} a_{3j}x_j^k\right) \\
&\vdots \\
x_n^{k+1} &= \frac{1}{a_{nn}}\left(b_n - \sum_{j=1}^{n-1} a_{nj}x_j^k\right)
\end{aligned} \tag{4.23}$$

qui consiste à isoler le coefficient de la diagonale de chaque ligne du système. C'est la *méthode de Jacobi*. Si l'un des coefficients diagonaux est nul, il est parfois possible de permuter certaines lignes pour éviter cette situation. Plus généralement, on écrit:

$$x_i^{k+1} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1,j\neq i}^{n} a_{ij}x_j^k\right) \tag{4.24}$$

Exemple 4.5
Soit le système:

$$\begin{aligned}
3x_1 + x_2 - x_3 &= 2 \\
x_1 + 5x_2 + 2x_3 &= 17 \\
2x_1 - x_2 - 6x_3 &= -18
\end{aligned}$$

La méthode de Jacobi s'écrit dans ce cas:

$$x_1^{k+1} = \frac{1}{3}\left(2 - x_2^k + x_3^k\right)$$

$$x_2^{k+1} = \frac{1}{5}\left(17 - x_1^k - 2x_3^k\right)$$

$$x_3^{k+1} = \frac{1}{-6}\left(-18 - 2x_1^k + x_2^k\right)$$

À partir de $[0\ \ 0\ \ 0]^T$, on trouve d'abord:

$$x_1^1 = \tfrac{1}{3}(2 - 0 + 0) = \tfrac{2}{3}$$

$$x_2^1 = \tfrac{1}{5}(17 - 0 - 0) = \tfrac{17}{5}$$

$$x_3^1 = \tfrac{1}{-6}(-18 - 0 + 0) = \tfrac{-18}{-6} = 3$$

La deuxième itération donne:

$$x_1^1 = \tfrac{1}{3}\left(2 - \tfrac{17}{5} + 3\right) = \tfrac{8}{15}$$

$$x_2^1 = \tfrac{1}{5}\left(17 - \tfrac{2}{3} - 2(3)\right) = \tfrac{31}{15}$$

$$x_3^1 = \tfrac{1}{-6}\left(-18 - 2(\tfrac{2}{3}) + \tfrac{17}{5}\right) = 2{,}655\,556$$

On finit par remplir le tableau suivant.

k	x_1^k	x_2^k	x_3^k
0	0,000 000	0,000 000	0,000 000
1	0,666 667	3,400 000	3,000 000
2	0,533 333	2,066 667	2,655 556
3	0,862 963	2,231 111	2,833 333
4	0,867 407	2,094 074	2,915 802
5	0,940 576	2,060 198	2,940 123
6	0,959 975	2,035 835	2,970 159
7	0,978 108	2,019 941	2,980 686
8	0,986 915	2,012 104	2,989 379
9	0,992 425	2,006 865	2,993 621
10	0,995 585	2,004 067	2,996 331

Les valeurs convergent vers la solution $[1\ \ 2\ \ 3]^T$. La convergence est cependant assez lente.

• • • • •

Exemple 4.6

La méthode de Jacobi ne peut pas s'appliquer immédiatement au système:

$$\begin{aligned} 0x_1 + 3x_2 + x_3 &= 7 \\ 5x_1 + x_2 - 2x_3 &= 15 \\ 3x_1 - 4x_2 + 8x_3 &= 9 \end{aligned}$$

puisque l'un des coefficients diagonaux est nul ($a_{11} = 0$). On remédie à cette situation en faisant pivoter les deux premières lignes, par exemple. On doit donc résoudre le système:

$$\begin{aligned} 5x_1 + x_2 - 2x_3 &= 15 \\ 0x_1 + 3x_2 + x_3 &= 7 \\ 3x_1 - 4x_2 + 8x_3 &= 9 \end{aligned}$$

pour lequel la méthode de Jacobi donne l'algorithme:

$$x_1^{k+1} = \frac{1}{5}\left(15 - x_2^k + 2x_3^k\right)$$

$$x_2^{k+1} = \frac{1}{3}\left(7 - x_3^k\right)$$

$$x_3^{k+1} = \frac{1}{8}\left(9 - 3x_1^k + 4x_2^k\right)$$

On obtient ainsi, en partant de la solution initiale $[0 \;\; 0 \;\; 0]^T$, les itérations suivantes.

k	x_1^k	x_2^k	x_3^k
1	3,000 000	2,333 333	1,125 000
2	2,983 333	1,958 333	1,166 667
3	3,075 000	1,944 444	0,985 417
4	3,005 278	2,004 861	0,944 097
5	2,976 667	2,018 634	1,000 451
6	2,996 454	1,999 850	1,018 067
7	3,007 257	1,993 978	1,001 255
8	3,001 706	1,999 582	0,994 268
9	2,997 791	2,001 911	0,999 151
10	2,999 278	2,000 283	1,001 784
11	3,000 657	1,999 405	1,000 412
12	3,000 284	1,999 863	0,999 456
13	2,999 810	2,000 181	0,999 825
14	2,999 894	2,000 058	1,000 162
15	3,000 053	1,999 946	1,000 069

Il y a donc convergence vers $[3\ \ 2\ \ 1]^T$.

●●●●●

Il est facile de montrer que la méthode de Jacobi peut s'écrire sous forme matricielle. Cette forme matricielle servira uniquement pour l'analyse de convergence de la méthode. Pour obtenir cette représentation matricielle, on doit d'abord effectuer une décomposition de la matrice A sous la forme:

$$A = D + T_i + T_s \tag{4.25}$$

où

$$D = \begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

$$T_i = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ a_{21} & 0 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & 0 \end{bmatrix}$$

$$T_s = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n} \\ 0 & 0 & 0 & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Ce procédé consiste donc à isoler de la matrice A la diagonale et les matrices triangulaires inférieures et supérieures. Le système linéaire:

$$A\vec{x} = \vec{b}$$

devient:

$$(D + T_i + T_s)\vec{x} = \vec{b}$$

ou encore

$$D\vec{x} = -(T_i + T_s)\vec{x} + \vec{b}$$

et enfin

$$\vec{x} = -D^{-1}(T_i + T_s)\vec{x} + D^{-1}\vec{b} = T_J\vec{x} + \vec{c}_J$$

où on a posé $T_J = -D^{-1}(T_i + T_s)$ et $\vec{c}_J = D^{-1}\vec{b}$. Il est alors clair que l'algorithme 4.24 peut aussi s'écrire:

$$\vec{x}^{k+1} = T_J\vec{x}^k + \vec{c}_J$$

et que la méthode de Jacobi n'est rien d'autre qu'un cas particulier de méthodes de points fixes en dimension n et de l'équation 4.10. Il suffit en effet de choisir:

$$\vec{g}(\vec{x}) = T_J\vec{x} + \vec{c}_J$$

Comme nous l'avons vu, la convergence des méthodes de points fixes en dimension n dépend du rayon spectral de la matrice jacobienne associée à $\vec{g}(\vec{x})$, qui est simplement (voir les exercices de fin de chapitre):

$$J = T_J = -(D^{-1}(T_i + T_s))$$

La méthode de Jacobi convergera donc si le rayon spectral de T_J ($\rho(T_J)$) est inférieur à 1. Comme nous l'avons déjà mentionné, le calcul du rayon spectral d'une matrice est un problème difficile, surtout si la matrice est de grande taille. Il existe cependant un type de matrices qui vérifie automatiquement la condition de convergence de la méthode de Jacobi.

Définition 4.10
Une matrice A est dite *à diagonale strictement dominante* si:

$$|a_{ii}| > \sum_{j=1, j \neq i}^{n} |a_{ij}| \quad \forall i$$

Cette définition signifie que le terme diagonal a_{ii} de la matrice A est nettement dominant puisque sa valeur absolue est plus grande que la somme des valeurs absolues de tous les autres termes de la ligne.

Pour démontrer que les matrices à diagonale dominante vérifient la condition $\rho(T_J) < 1$, un résultat intermédiaire est nécessaire.

Lemme 4.1
Le rayon spectral d'une matrice A vérifie:

$$\rho(A) \leq ||A|| \tag{4.26}$$

et ce quelle que soit la norme matricielle utilisée.

Démonstration:
Si λ est une valeur propre de la matrice A et $\vec{x}$ est un vecteur propre associé, on a:

$$A\vec{x} = \lambda\vec{x}$$

En prenant la norme de chaque côté (ce qui suppose l'utilisation d'une norme vectorielle compatible avec la norme matricielle utilisée), on obtient:

$$|\lambda| \, ||\vec{x}|| = ||A\vec{x}|| \leq ||A|| \, ||\vec{x}||$$

ce qui entraîne immédiatement que:

$$|\lambda| \leq ||A||$$

Ceci étant vrai quelle que soit la valeur propre choisie, le résultat est démontré. □

Théorème 4.5

Si A est une matrice à diagonale strictement dominante, la méthode de Jacobi est convergente.

Démonstration:

Il suffit de montrer que $\rho(-D^{-1}(T_i + T_s)) < 1$. Si on utilise les normes vectorielle et matricielle compatibles $|| \, ||_\infty$, il est clair que:

$$|| - D^{-1}(T_i + T_s)||_\infty < 1$$

et le lemme précédent confirme que le rayon spectral de la matrice est aussi inférieur à 1. □

Exemple 4.7

Il est facile de s'assurer que les matrices des deux exemples de cette section sont à diagonale strictement dominante. Nous avons pu constater la convergence de la méthode de Jacobi dans ces deux cas, comme le prévoit le théorème précédent. Pour s'en convaincre davantage, on peut construire explicitement les matrices $-D^{-1}(T_i + T_s)$ de ces deux exemples et constater que leur norme $|| \, ||_\infty$ est inférieure à 1.

●●●●●

4.5.2 Méthode de Gauss-Seidel

La méthode de Gauss-Seidel est une variante améliorée de la méthode de Jacobi. Pour bien en comprendre le principe, il suffit de reconsidérer la méthode de Jacobi et de voir comment on pourrait l'améliorer. On sait que dans le cas général la méthode de Jacobi s'écrit:

$$x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{n-1} a_{ij} x_j^k \right)$$

qui peut aussi s'exprimer:

$$x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^k - \sum_{j=i+1}^{n} a_{ij} x_j^k \right) \qquad (4.27)$$

La méthode de Gauss-Seidel est fondée sur la simple constatation selon laquelle le calcul de x_2^{k+1} nécessite l'utilisation de $x_1^k, x_3^k, \cdots, x_n^k$ provenant de l'itération précédente. Or, à l'itération $k+1$, au moment du calcul de x_2^{k+1}, on possède déjà une meilleure approximation de x_1 que x_i^k, à savoir x_1^{k+1}. De même, au moment du calcul de x_3^{k+1}, on peut utiliser x_1^{k+1} et x_2^{k+1}. Plus généralement, pour le calcul de x_i^{k+1}, on peut utiliser $x_1^{k+1}, x_2^{k+1}, \cdots, x_{i-1}^{k+1}$ déjà calculés et les $x_{i+1}^k, x_{i+2}^k, \cdots, x_n^k$ de l'itération précédente. Cela revient à écrire:

$$x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^k \right) \tag{4.28}$$

Suivant la notation introduite pour la méthode de Jacobi, la méthode de Gauss-Seidel s'écrit sous forme matricielle:

$$\vec{x}^{k+1} = D^{-1}(\vec{b} - T_i \vec{x}^{k+1} - T_s \vec{x}^k)$$

ou encore

$$(T_i + D)\vec{x}^{k+1} = \vec{b} - T_s \vec{x}^k$$

et enfin

$$\vec{x}^{k+1} = -(T_i + D)^{-1} T_s \vec{x}^k + (T_i + D)^{-1} \vec{b} = T_{GS} \vec{x}^k + \vec{c}_{GS}$$

Exemple 4.8
Soit le système linéaire:

$$\begin{aligned} 3x_1 + x_2 - x_3 &= 2 \\ x_1 + 5x_2 + 2x_3 &= 17 \\ 2x_1 - x_2 - 6x_3 &= -18 \end{aligned}$$

La méthode de Gauss-Seidel s'écrit dans ce cas:

$$x_1^{k+1} = \tfrac{1}{3}\left(2 - x_2^k + x_3^k\right)$$

$$x_2^{k+1} = \tfrac{1}{5}\left(17 - x_1^{k+1} - 2x_3^k\right)$$

$$x_3^{k+1} = \tfrac{1}{-6}\left(-18 - 2x_1^{k+1} + x_2^{k+1}\right)$$

Notons la différence avec la méthode de Jacobi au niveau des indices. Partant de $[0 \;\; 0 \;\; 0]^T$, on trouve d'abord:

$$x_1^1 = \tfrac{1}{3}(2 - 0 + 0) = \tfrac{2}{3}$$

$$x_2^1 = \tfrac{1}{5}\left(17 - \tfrac{2}{3} - 0\right) = \tfrac{49}{15}$$

$$x_3^1 = \tfrac{1}{-6}\left(-18 - 2(\tfrac{2}{3}) + \tfrac{49}{15}\right) = \tfrac{241}{90}$$

tandis qu'à la deuxième itération on trouve:

$$x_1^1 = \tfrac{1}{3}\left(2 - \tfrac{49}{15} + \tfrac{241}{90}\right) = 0{,}470\,3704$$

$$x_2^1 = \tfrac{1}{5}\left(17 - 0{,}470\,3704 - 2(\tfrac{241}{90})\right) = 2{,}234\,815$$

$$x_3^1 = \tfrac{1}{-6}\left(-18 - 2(0{,}470\,3704) + 2{,}234\,815\right) = 2{,}784\,321$$

ainsi que les itérations suivantes.

k	x_1^k	x_2^k	x_3^k
1	0,666 6667	3,266 667	2,677 778
2	0,470 3704	2,234 815	2,784 321
3	0,849 8354	2,116 305	2,930 561
4	0,938 0855	2,040 158	2,972 669
5	0,977 5034	2,015 432	2,989 929
6	0,991 4991	2,005 729	2,996 212
7	0,996 8277	2,002 150	2,998 584
8	0,998 8115	2,000 804	2,999 470
9	0,999 5553	2,000 301	2,999 802
10	0,999 8335	2,000 113	2,999 926

On constate que, pour un même nombre d'itérations, la solution approximative obtenue par la méthode de Gauss-Seidel est plus précise. La méthode de Gauss-Seidel converge généralement plus vite que la méthode de Jacobi, mais pas toujours.

● ● ● ● ●

Remarque 4.13

Les méthodes itératives utilisées pour résoudre un système linéaire de la forme:

$$A\vec{x} = \vec{b}$$

s'écrivent souvent sous la forme:

$$\vec{x}^{k+1} = T\vec{x}^k + \vec{c}$$

où la matrice T et le vecteur $\vec{c}$ dépendent de la méthode en cause. Les méthodes de Jacobi et de Gauss-Seidel ne sont que deux cas particuliers. La convergence de la méthode retenue n'est possible que si $\rho(T) < 1$. *Cette condition porte sur la matrice T et non sur la matrice A.* La convergence de la méthode est d'autant plus rapide que le rayon spectral de T est petit. En effet, on peut démontrer que:

$$||\vec{x}^{k+1} - \vec{x}|| \simeq (\rho(T))^{k+1}||\vec{x}^0 - \vec{x}||$$

Cette expression s'apparente à celle du taux de convergence $g'(r)$ dans le cas unidimensionnel. □

Nous terminons cette section par un résultat que nous ne démontrons pas, mais qui assure la convergence de la méthode de Gauss-Seidel dans le cas de matrices à diagonale strictement dominante.

Théorème 4.6

Si la matrice A est à diagonale strictement dominante, le rayon spectral de T_{GS} est inférieur à 1 et la méthode de Gauss-Seidel converge, et ce quelle que soit la solution initiale $\vec{x}^0$. □

Les matrices à diagonale strictement dominante sont fréquentes dans les applications. Malheureusement, on rencontre également en pratique beaucoup de matrices qui ne vérifient pas cette propriété. Des recherches très actives sont en cours dans le but de développer de nouvelles méthodes itératives qui s'appliqueraient à une vaste gamme de matrices.

4.6 Exercices

1. Montrer que tout point fixe de $g(x)$ est un point fixe de $g(g(x))$. L'inverse est-il vrai?

2. Montrer que les points x_1 et x_2 de l'équation 4.4 satisfont $x_1 = g(x_2)$ et $x_2 = g(x_1)$.

3. Déterminer le polynôme caractéristique, les valeurs propres et le rayon spectral des matrices suivantes. Déterminer également si ces matrices sont convergentes.

$$\text{a)} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \quad \text{b)} \begin{bmatrix} \frac{1}{2} & 1 & 2 \\ 0 & \frac{1}{3} & 3 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}$$

4. La matrice:

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

a un rayon spectral de 3, qui est donc supérieur à 1. Or, on peut démontrer directement que la méthode de Jacobi converge quel que soit le vecteur $\vec{x}_0$ initial. Est-ce une contradiction?

5. a) Obtenir le ou les points fixes de l'application:

$$\begin{aligned} x_1 &= \sqrt{2 - x_2^2} \\ x_2 &= \sqrt{x_1} \end{aligned}$$

b) Vérifier si ce ou ces points fixes sont attractifs.

c) Effectuer 5 itérations de la méthode des points fixes à partir de (0,0).

6. a) Montrer que l'ensemble $\{x_1, x_2\}$ (attracteur de période 2) est attractif si:

$$|g'(x_1)g'(x_2)| < 1$$

b) En déduire une condition pour qu'un attracteur de période n soit attractif.

7. Trouver un attracteur de période 2 pour la fonction $g(x) = x^2 - 1$. Vérifier s'il est attractif.

8. Montrer que la fonction:

$$T(x) = \begin{cases} 2x & \text{pour} \quad 0 \leq x \leq 1/2 \\ 2 - 2x & \text{pour} \quad 1/2 \leq x \leq 1 \end{cases}$$

possède les attracteurs de période 3 $\{2/7, 4/7, 6/7\}$ ainsi que $\{2/9, 4/9, 8/9\}$. Vérifier s'ils sont attractifs ou répulsifs.

9. Montrer que la matrice jacobienne associée à la méthode de points fixes:

$$\vec{g}(\vec{x}) = T\vec{x} + \vec{c}$$

où T est une matrice de dimension n, est tout simplement la matrice T elle-même. Que conclure quant à la convergence de l'algorithme:

$$\vec{x}^{k+1} = T\vec{x}^{k} + \vec{c}$$

10. Construire une matrice de dimension 3 à diagonale strictement dominante. Pour cette matrice, construire explicitement la matrice T_J et vérifier si $||T_J||_\infty < 1$.

11. Soit le système linéaire suivant:

$$\begin{array}{lrcr} E_1: & 2x_1 - x_2 + 10x_3 & = & -11 \\ E_2: & 3x_2 - x_3 + 8x_4 & = & -11 \\ E_3: & 10x_1 - x_2 + 2x_3 & = & 6 \\ E_4: & -x_1 + 11x_2 - x_3 + 3x_4 & = & 25 \end{array}$$

a) Montrer que les méthodes de Jacobi et de Gauss-Seidel ne convergent pas lorsqu'on isole simplement les x_i de l'équation E_i.

b) Réordonner les équations de façon à assurer la convergence des deux méthodes.

12. Résoudre le système:

$$\begin{array}{rcr} 9x_1 - 2x_2 + x_3 & = & 13 \\ -x_1 + 5x_2 - x_3 & = & 9 \\ x_1 - 2x_2 + 9x_3 & = & -11 \end{array}$$

à l'aide des méthodes de Jacobi et de Gauss-Seidel à partir de $\vec{x}_0 = [0 \;\; 0 \;\; 0]^T$ (faire les 5 premières itérations seulement).

Chapitre 5

Interpolation

5.1 Introduction

Ce chapitre ainsi que le chapitre suivant qui porte sur la dérivation et l'intégration numériques sont très étroitement reliés puisqu'ils tendent à répondre à diverses facettes d'un même problème. Ce problème est le suivant: à partir d'une fonction $f(x)$ connue seulement en $(n+1)$ points de la forme $((x_i, f(x_i))$ pour $i = 0, 1, 2, \cdots, n)$, peut-on construire une approximation de $f(x)$, et ce pour tout x? Les points $((x_i, f(x_i))$ pour $i = 0, 1, 2, \cdots, n)$ sont appelés *points de collocation* ou *points d'interpolation* et peuvent provenir de données expérimentales ou d'une table. En d'autres termes, si on ne connaît que les points de collocation $(x_i, f(x_i))$ d'une fonction, peut-on obtenir une approximation de $f(x)$ pour une valeur de x différente des x_i? La figure 5.1 résume la situation.

Sur la base des mêmes hypothèses, nous verrons, au chapitre suivant, comment évaluer les dérivées $f'(x), f''(x) \cdots$ de même que:

$$\int_{x_0}^{x_n} f(x)dx$$

Il s'agit d'un *problème d'interpolation*, dont la solution est relativement simple. Il suffit de construire un polynôme de degré suffisamment élevé dont la courbe passe par les points de collocation. On parle alors du *polynôme de collocation* ou *polynôme d'interpolation*. Pour obtenir une approximation des dérivées ou de l'intégrale, il suffit de dériver ou d'intégrer le polynôme de collocation. Il y a cependant des éléments fondamentaux qu'il est important d'étudier. En premier lieu, il convient de rappeler certains résultats cruciaux relatifs aux polynômes, que nous ne démontrons pas.

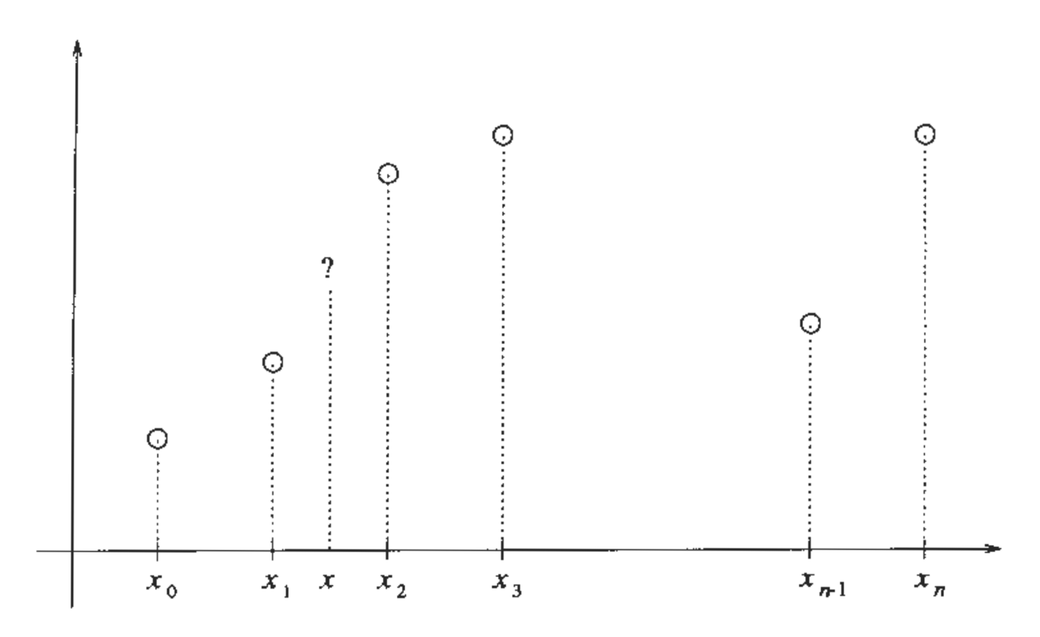


Figure 5.1: Problème d'interpolation

Théorème 5.1

Un polynôme de degré n dont la forme générale est:

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n \quad (a_n \neq 0) \tag{5.1}$$

possède très exactement n racines qui peuvent être réelles ou complexes conjuguées. (On sait que r est une racine de $p_n(x)$ si $p_n(r) = 0$.) □

Corollaire 5.1

*Par $(n+1)$ points de collocation $((x_i, f(x_i))$ pour $i = 0, 1, 2, \cdots, n)$, on ne peut faire correspondre qu'*un et un seul *polynôme de degré n.*

Démonstration:

On procède par l'absurde et on suppose l'existence de 2 polynômes de degré n, notés $p(x)$ et $q(x)$, et qui passent tous les deux par les $(n+1)$ points de collocation donnés. On considère ensuite la différence:

$$P(x) = p(x) - q(x)$$

qui est également un polynôme de degré au plus n. Ce polynôme vérifie:

$$P(x_i) = p(x_i) - q(x_i) = f(x_i) - f(x_i) = 0$$

et ce pour i allant de 0 à n. Le polynôme $P(x)$ posséderait donc $(n+1)$ racines, ce qui est impossible en vertu du théorème précédent.□

Remarque 5.1
Le raisonnement précédent établit en fait l'unicité du polynôme d'interpolation passant par $n+1$ points donnés. Il reste à en assurer l'existence, ce que nous ferons tout simplement en le construisant au moyen de méthodes diverses qui feront l'objet des prochaines sections.□

5.2 Matrice de Vandermonde

Le problème d'interpolation consiste donc à déterminer l'unique polynôme de degré n passant par les $(n+1)$ points de collocation $((x_i, f(x_i))$ pour $i = 0, 1, 2, 3, \cdots, n)$. Selon le théorème précédent, il ne saurait y en avoir deux. Il reste maintenant à le construire de la manière la plus efficace et la plus générale possible. Une première tentative consiste à déterminer les inconnues a_i du polynôme 5.1 en vérifiant directement les $(n+1)$ équations de collocation:

$$p_n(x_i) = f(x_i) \quad \text{pour } i = 0, 1, 2, \cdots, n$$

ou encore

$$a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 + \cdots + a_n x_i^n = f(x_i)$$

qui est un système linéaire de $(n+1)$ équations en $(n+1)$ inconnues. Ce système s'écrit sous forme matricielle:

$$\begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & x_1^3 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & x_2^3 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix} \tag{5.2}$$

Remarque 5.2
La matrice de ce système linéaire porte le nom de *matrice de Vandermonde*. On peut montrer que le conditionnement de cette matrice augmente fortement avec la taille $(n+1)$ du système. De plus, comme le révèlent les sections qui suivent, il n'est pas nécessaire de résoudre un système linéaire pour calculer un polynôme d'interpolation. *Cette méthode est donc rarement utilisée.*□

Exemple 5.1
On doit calculer le polynôme passant par les points $(0\,,\,1), (1\,,\,2), (2\,,\,9)$ et $(3\,,\,28)$. Étant donné ces 4 points, le polynôme recherché est tout au plus de degré 3. Ses coefficients a_i sont solution de:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 9 \\ 28 \end{bmatrix}$$

dont la solution (obtenue par décomposition LU) est $[1\ 0\ 0\ 1]^T$. Le polynôme recherché est donc:

$$p_3(x) = 1 + x^3$$

● ● ● ● ●

Les sections qui suivent proposent des avenues différentes et plus efficaces pour calculer le polynôme de collocation.

5.3 Interpolation de Lagrange

L'interpolation de Lagrange est une façon simple et systématique de construire un polynôme de collocation. Étant donné $(n+1)$ points $((x_i, f(x_i))$ pour $i = 0, 1, 2, \cdots, n)$, on suppose un instant que l'on sait construire $(n+1)$ polynômes $L_i(x)$ de degré n et satisfaisant les conditions suivantes:

$$\begin{array}{rcll} L_i(x_i) & = & 1 & \forall i \\ L_i(x_j) & = & 0 & \forall j \neq i \end{array} \tag{5.3}$$

Cela signifie que le polynôme $L_i(x)$ de degré n prend la valeur 1 en x_i et s'annule à tous les autres points de collocation. Nous verrons comment construire les $L_i(x)$ un peu plus loin. Dans ces conditions, la fonction $L(x)$ définie par:

$$L(x) = \sum_{i=0}^{n} f(x_i)L_i(x)$$

est un polynôme de degré n, car chacun des $L_i(x)$ est de degré n. De plus, ce polynôme passe par les $(n+1)$ points de collocation et est donc le polynôme recherché. En effet, il est facile de montrer que selon les conditions 5.3:

$$\begin{aligned} L(x_j) &= f(x_j)L_j(x_j) + \sum_{i=0, i\neq j}^{n} f(x_i)L_i(x_j) \\ &= f(x_j) + 0 = f(x_j) \ \ \forall j \end{aligned}$$

Le polynôme $L(x)$ passe donc par tous les points de collocation. Puisque ce polynôme est unique, $L(x)$ est bien le polynôme recherché. Il reste à construire les fonctions $L_i(x)$. Suivons une démarche progressive.

Polynômes de degré 1

Il s'agit de déterminer le polynôme de degré 1 dont la courbe (une droite) passe par les deux points $(x_0, f(x_0))$ et $(x_1, f(x_1))$. On doit donc construire deux polynômes $L_0(x)$ et $L_1(x)$ de degré 1 qui vérifient:

$$\begin{aligned} L_0(x_0) &= 1 \\ L_0(x_1) &= 0 \end{aligned}$$

$$\begin{aligned} L_1(x_0) &= 0 \\ L_1(x_0) &= 1 \end{aligned}$$

Le polynôme $L_0(x)$ doit s'annuler en $x = x_1$. On pense immédiatement au polynôme:

$$(x - x_1)$$

qui s'annule en $x = x_1$, mais qui vaut $(x_0 - x_1)$ en $x = x_0$. Pour s'assurer d'une valeur 1 en $x = x_0$, il suffit d'effectuer la division appropriée afin d'obtenir:

$$L_0(x) = \frac{(x - x_1)}{(x_0 - x_1)}$$

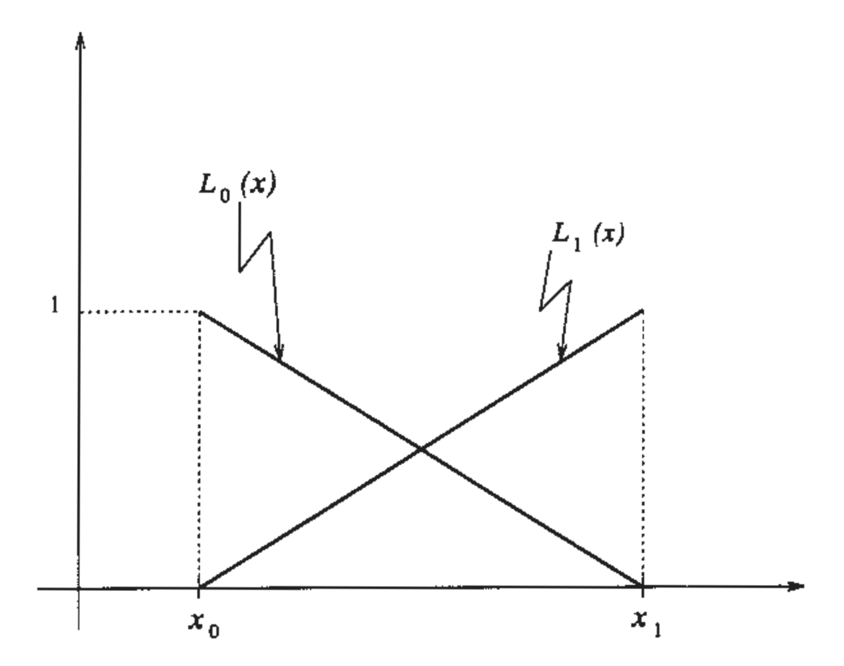


Figure 5.2: Polynômes de Lagrange de degré 1: $L_0(x)$ et $L_1(x)$

Un raisonnement similaire pour $L_1(x)$ donne:

$$L_1(x) = \frac{(x - x_0)}{(x_1 - x_0)}$$

Ces deux fonctions sont illustrées à la figure 5.2. Le polynôme de degré 1 est donc:

$$p_1(x) = f(x_0)L_0(x) + f(x_1)L_1(x)$$

Exemple 5.2

L'équation de la droite passant par les points (2 , 3) et (5 , −6) est:

$$3\frac{(x-5)}{(2-5)} + (-6)\frac{(x-2)}{(5-2)} = -(x-5) - 2(x-2) = -3x + 9$$

●●●●●

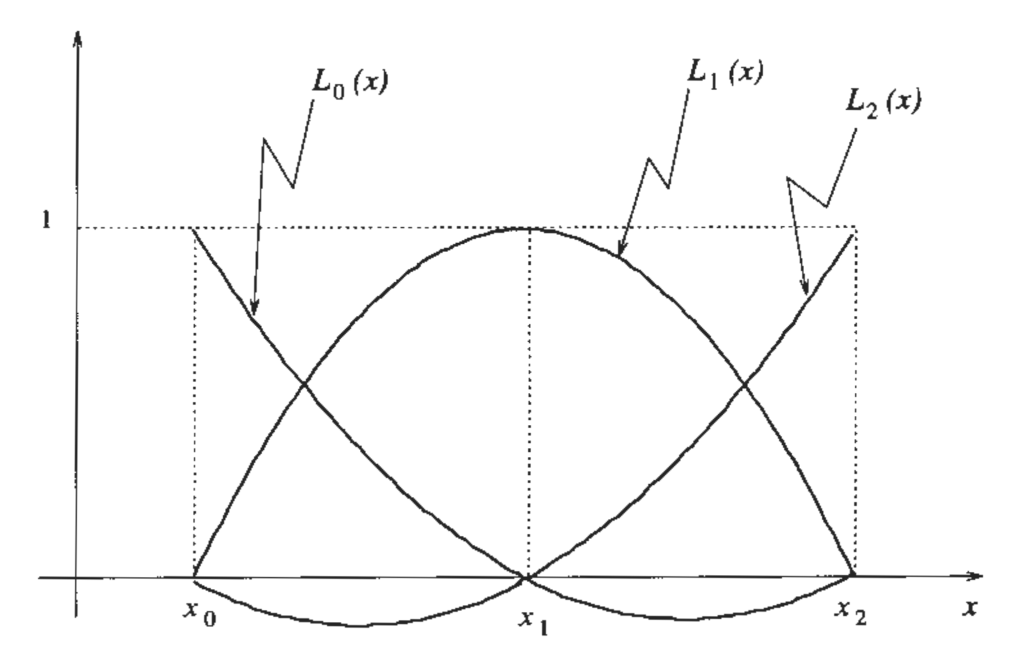


Figure 5.3: Polynômes de Lagrange de degré 2: $L_0(x), L_1(x)$ et $L_2(x)$

Polynômes de degré 2

Si on cherche le polynôme de degré 2 passant par les trois points $(x_0, f(x_0)), (x_1, f(x_1))$ et $(x_2, f(x_2))$, on doit construire trois fonctions $L_i(x)$. Le raisonnement est toujours le même. La fonction $L_0(x)$ s'annule cette fois en $x = x_1$ et en $x = x_2$. On doit forcément avoir un coefficient de la forme:

$$(x - x_1)(x - x_2)$$

qui vaut $(x_0-x_1)(x_0-x_2)$ en $x = x_0$. Pour satisfaire la condition $L_0(x_0) = 1$, il suffit alors de diviser le coefficient par cette valeur et de poser:

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}$$

Cette fonction vaut bien 1 en x_0 et 0 en x_1 et x_2. De la même manière, on obtient les fonctions $L_1(x)$ et $L_2(x)$ définies par:

$$L_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} \quad \text{et} \quad L_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}$$

Ces trois fonctions sont à leur tour illustrées à la figure 5.3.

Exemple 5.3
La parabole passant par les points $(1\,,\,2),(3\,,\,7),(4\,,\,-1)$ est donnée par:

$$\begin{aligned} p_2(x) &= 2\frac{(x-3)(x-4)}{(1-3)(1-4)} + 7\frac{(x-1)(x-4)}{(3-1)(3-4)} + (-1)\frac{(x-1)(x-3)}{(4-1)(4-3)} \\ &= \frac{(x-3)(x-4)}{3} - \frac{7(x-1)(x-4)}{2} - \frac{(x-1)(x-3)}{3} \end{aligned}$$

●●●●●

Polynômes de degré n

On analyse le cas général de la même façon. La fonction $L_0(x)$ doit s'annuler en $x = x_1, x_2, x_3, \cdots, x_n$. Il faut donc introduire la fonction:

$$(x-x_1)(x-x_2)(x-x_3)\cdots(x-x_n)$$

qui vaut:

$$(x_0-x_1)(x_0-x_2)(x_0-x_3)\cdots(x_0-x_n)$$

en $x = x_0$. On a alors, après division:

$$L_0(x) = \frac{(x-x_1)(x-x_2)(x-x_3)\cdots(x-x_n)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)\cdots(x_0-x_n)}$$

On remarque qu'il y a n facteurs de la forme $(x - x_i)$ dans cette expression et qu'il s'agit bien d'un polynôme de degré n. Pour la fonction $L_1(x)$, on pose:

$$L_1(x) = \frac{(x-x_0)(x-x_2)(x-x_3)\cdots(x-x_n)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)\cdots(x_1-x_n)}$$

On note l'absence du terme $(x - x_1)$. L'expression générale pour la fonction $L_i(x)$ est donc:

$$L_i(x) = \frac{(x-x_0)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)} \qquad (5.4)$$

où cette fois seul le facteur $(x - x_i)$ est absent. $L_i(x)$ est donc un polynôme de degré n qui vaut 1 en $x = x_i$ et qui s'annule à tous les autres points de collocation. On peut maintenant résumer la situation.

Théorème 5.2

Étant donné $(n+1)$ points d'interpolation $((x_i, f(x_i))$ pour $i = 0, 1, \cdots, n)$, l'unique polynôme d'interpolation de degré n passant par tous ces points peut s'écrire:

$$p_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x) \tag{5.5}$$

où les $(n+1)$ fonctions $L_i(x)$ sont définies par la relation 5.4. C'est la *formule de Lagrange.* □

Exemple 5.4

Reprenons les points $(0\,,\ 1), (1\,,\ 2), (2\,,\ 9)$ et $(3\,,\ 28)$, pour lesquels nous avons obtenu le polynôme $p_3(x) = x^3 + 1$ à l'aide de la matrice de Vandermonde. L'interpolation de Lagrange donne dans ce cas:

$$\begin{aligned} p_3(x) &= 1\frac{(x-1)(x-2)(x-3)}{(0-1)(0-2)(0-3)} + 2\frac{(x-0)(x-2)(x-3)}{(1-0)(1-2)(1-3)} \\ &\quad +9\frac{(x-0)(x-1)(x-3)}{(2-0)(2-1)(2-3)} + 28\frac{(x-0)(x-1)(x-2)}{(3-0)(3-1)(3-2)} \end{aligned}$$

c'est-à-dire:

$$\begin{aligned} p_3(x) &= -\frac{(x-1)(x-2)(x-3)}{6} + x(x-2)(x-3) \\ &\quad -9\frac{x(x-1)(x-3)}{2} + 14\frac{x(x-1)(x-2)}{3} \end{aligned}$$

qui est l'expression du polynôme de degré 3 passant par les 4 points donnés. Cette expression n'est autre que $p_3(x) = x^3 + 1$. Il n'y a qu'à en faire le développement pour s'en assurer. Cela n'est pas surprenant, puisque l'on sait qu'il n'existe qu'un seul polynôme de degré 3 passant par 4 points donnés. L'interpolation de Lagrange ne fait qu'exprimer le même polynôme différemment.

Enfin, le polynôme calculé permet d'obtenir une approximation de la fonction inconnue $f(x)$ partout dans l'intervalle contenant les points de collocation, c'est-à-dire $[0, 3]$. Ainsi, on a:

$$f(2{,}5) \simeq p_3(2{,}5) = 16{,}625$$

avec une précision qui sera discutée plus loin lorsque nous aborderons la question de l'erreur d'interpolation.

● ● ● ● ●

Remarque 5.3
La méthode d'interpolation de Lagrange présente un inconvénient majeur: elle n'est pas récursive. En effet, si on souhaite passer d'un polynôme de degré n à un polynôme de degré $(n + 1)$ (en ajoutant un point de collocation), on doit reprendre tout le processus à zéro. Dans l'exemple précédent, si on souhaite obtenir le polynôme de degré 4 correspondant aux points $(0\,,\ 1), (1\,,\ 2), (2\,,\ 9), (3\,,\ 28)$ et $(5\,,\ 54)$, on ne peut d'aucune façon récupérer le polynôme de degré 3 déjà calculé et le modifier simplement pour obtenir $p_4(x)$. C'est en revanche ce que permet la méthode d'interpolation de Newton. □

5.4 Polynôme de Newton

Lorsqu'on écrit l'expression générale d'un polynôme, on pense immédiatement à la forme 5.1, qui est la plus utilisée. Il en existe cependant d'autres qui sont plus appropriées au cas de l'interpolation, par exemple:

$$\begin{aligned} p_n(x) \quad = \quad & a_0 \\ & + a_1(x - x_0) \\ & + a_2(x - x_0)(x - x_1) \\ & + a_3(x - x_0)(x - x_1)(x - x_2) \\ & \vdots \\ & + a_{n-1}(x - x_0)(x - x_1)(x - x_2)\cdots(x - x_{n-2}) \\ & + a_n(x - x_0)(x - x_1)(x - x_2)\cdots(x - x_{n-1}) \end{aligned} \tag{5.6}$$

On remarque que le coefficient de a_n comporte n monômes de la forme $(x-x_i)$ et qu'en conséquence le polynôme 5.6 est de degré n.

L'aspect intéressant de cette formule apparaît lorsqu'on essaie de déterminer les $(n+1)$ coefficients a_i de telle sorte que $p_n(x)$ passe par les $(n+1)$ points de collocation $(x_i, f(x_i))$ pour $i = 0, 1, 2, \cdots, n$). On doit donc s'assurer que:

$$p_n(x_i) = f(x_i) \quad \text{pour} \quad i = 0, 1, 2, \cdots, n$$

Les coefficients de la forme 5.6 s'annulent tous en $x = x_0$, sauf le premier. On peut ainsi montrer que:

$$p_n(x_0) = a_0 = f(x_0)$$

Le premier coefficient est donc:

$$a_0 = f(x_0) \tag{5.7}$$

On doit ensuite s'assurer que $p_n(x_1) = f(x_1)$, c'est-à-dire:

$$p_n(x_1) = a_0 + a_1(x_1 - x_0) = f(x_0) + a_1(x_1 - x_0) = f(x_1)$$

ce qui permet d'isoler a_1 pour obtenir:

$$a_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

Définition 5.1
On définit les *premières différences divisées* de la fonction $f(x)$ par:

$$f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \tag{5.8}$$

Ainsi, le coefficient a_1 peut s'écrire:

$$a_1 = f[x_0, x_1] \tag{5.9}$$

Remarque 5.4
Il est facile de démontrer que le polynôme de degré 1:

$$p_1(x) = f(x_0) + f[x_0, x_1](x - x_0)$$

qu'on obtient en ne considérant que les deux premiers coefficients de 5.6 et les expressions 5.7 et 5.9, passe par les points $(x_0, f(x_0))$ et $(x_1, f(x_1))$. Il représente donc l'unique polynôme de collocation de degré 1 passant par ces deux points. □

Le troisième coefficient (a_2) est à son tour déterminé par:

$$p_n(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) = f(x_2)$$

ou encore

$$p_n(x_2) = f(x_0) + f[x_0, x_1](x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) = f(x_2)$$

En isolant a_2, on obtient:

$$\begin{aligned}
a_2 &= \frac{1}{(x_2 - x_0)(x_2 - x_1)} \left(f(x_2) - f(x_0) - f[x_0, x_1](x_2 - x_0)\right) \\
&= \frac{1}{(x_2 - x_0)} \left(\frac{f(x_2) - f(x_0)}{(x_2 - x_1)} - f[x_0, x_1]\frac{(x_2 - x_0)}{(x_2 - x_1)}\right) \\
&= \frac{1}{(x_2 - x_0)} \left(\frac{f(x_2) - f(x_1) + f(x_1) - f(x_0)}{(x_2 - x_1)} - f[x_0, x_1]\frac{(x_2 - x_0)}{(x_2 - x_1)}\right) \\
&= \frac{1}{(x_2 - x_0)} \left(\frac{f(x_2) - f(x_1)}{(x_2 - x_1)} + \frac{(f(x_1) - f(x_0))}{(x_1 - x_0)}\frac{(x_1 - x_0)}{(x_2 - x_1)}\right. \\
&\quad \left. - f[x_0, x_1]\frac{(x_2 - x_0)}{(x_2 - x_1)}\right) \\
&= \frac{1}{(x_2 - x_0)} \left(f[x_1, x_2] + f[x_0, x_1]\left(\frac{(x_1 - x_0)}{(x_2 - x_1)} - \frac{(x_2 - x_0)}{(x_2 - x_1)}\right)\right) \\
&= \frac{1}{(x_2 - x_0)} \left(f[x_1, x_2] - f[x_0, x_1]\right)
\end{aligned}$$

On en arrive donc à une expression qui fait intervenir une différence divisée de différences divisées.

Définition 5.2
Les *deuxièmes différences divisées* de la fonction $f(x)$ sont définies à partir des premières différences divisées par la relation:

$$f[x_i, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{(x_{i+2} - x_i)} \tag{5.10}$$

De même, les n^{es} différences divisées de la fonction $f(x)$ sont définies à partir des $(n-1)^{es}$ différences divisées de la façon suivante:

$$f[x_0, x_1, x_2, \cdots, x_n] = \frac{f[x_1, x_2, \cdots, x_n] - f[x_0, x_1, x_2, \cdots, x_{n-1}]}{(x_n - x_0)} \tag{5.11}$$

Notons que les toutes premières différences divisées de $f(x)$ (soit les 0es différences) sont tout simplement définies par $f(x_i)$.

Suivant cette notation, on a:

$$a_2 = f[x_0, x_1, x_2] \tag{5.12}$$

Remarque 5.5
Il est facile de démontrer que le polynôme:

$$p_2(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)$$

passe par les trois premiers points de collocation. On remarque de plus que ce polynôme de degré 2 s'obtient simplement par l'ajout d'un terme de degré 2 au polynôme $p_1(x)$ déjà calculé. En raison de cette propriété, cette méthode est dite *récursive*. □

On peut soupçonner à ce stade-ci que le coefficient a_3 est:

$$a_3 = f[x_0, x_1, x_2, x_3]$$

qui est une troisième différence divisée de $f(x)$. C'est effectivement le cas. Le théorème suivant résume la situation.

Théorème 5.3

L'unique polynôme de degré n passant par les $(n+1)$ points de collocation $((x_i, f(x_i))$ pour $i = 0, 1, 2, \cdots, n)$ peut s'écrire selon la *formule d'interpolation de Newton* 5.6 ou encore sous la forme récursive:

$$p_n(x) = p_{n-1}(x) + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}) \qquad (5.13)$$

Les coefficients de ce polynôme sont les différences divisées:

$$a_i = f[x_0, x_1, x_2, \cdots, x_i] \quad \text{pour } 0 \leq i \leq n \qquad \square \qquad (5.14)$$

Démonstration (facultative):

On démontre le résultat par induction. On a déjà établi le résultat pour $n = 1$ et $n = 2$. On suppose que ce résultat est vrai pour les polynômes de degré $(n-1)$. Il s'agit de montrer qu'il est également vrai pour les polynômes de degré n. Pour ce faire, on introduit les polynômes $p_{n-1}(x)$ et $q_{n-1}(x)$ de degré $(n-1)$ et passant respectivement par les points $((x_i, f(x_i))$ pour $i = 0, 1, 2, \cdots, n-1)$, et $((x_i, f(x_i))$ pour $i = 1, 2, 3, \cdots, n)$. On note immédiatement que ces deux polynômes passent respectivement par les n premiers et les n derniers points d'interpolation. Les coefficients a_i étant définis par la relation 5.14, on pose également:

$$b_i = f[x_1, x_2, \cdots, x_{i+1}] \quad \text{pour } 1 \leq i \leq n-1$$

Les a_i et les b_i sont les différences divisées relatives aux n premiers et aux n derniers points, respectivement. Suivant la définition des différences divisées, on observe que:

$$a_n = f[x_0, x_1, \cdots, x_n] = \frac{f[x_1, x_2, \cdots, x_n] - f[x_0, x_1, \cdots, x_{n-1}]}{x_n - x_0}$$

c'est-à-dire:

$$a_n = \frac{b_{n-1} - a_{n-1}}{x_n - x_0} \qquad (5.15)$$

L'hypothèse d'induction permet d'affirmer que:

$$p_{n-1}(x) = p_{n-2}(x) + a_{n-1}(x - x_0)(x - x_1)\cdots(x - x_{n-1}) \qquad (5.16)$$

et que:

$$q_{n-1}(x) = q_{n-2}(x) + b_{n-1}(x - x_1)(x - x_2)\cdots(x - x_n) \qquad (5.17)$$

La démonstration du théorème requiert de plus l'utilisation du lemme suivant.

Lemme 5.1
L'unique polynôme $p_n(x)$ *passant par les points* $((x_i, f(x_i))$ *pour* $i = 0, 1, 2, \cdots, n)$ *s'écrit:*

$$p_n(x) = \frac{(x_n - x)p_{n-1}(x) + (x - x_0)q_{n-1}(x)}{(x_n - x_0)} \tag{5.18}$$

Preuve du lemme:
Il suffit de s'assurer que le polynôme à droite de l'équation 5.18, qui est de degré n et noté $r_n(x)$, passe également par les points $((x_i, f(x_i))$ pour $i = 0, 1, 2, \cdots, n)$. Le résultat suivra par unicité du polynôme d'interpolation. Suivant la définition des polynômes $r_n(x)$, $p_{n-1}(x)$ et $q_{n-1}(x)$, on a:

$$r_n(x_0) = p_{n-1}(x_0) = f(x_0)$$

et à l'autre extrémité:

$$r_n(x_n) = q_{n-1}(x_n) = f(x_n)$$

Aux points intermédiaires, on a:

$$\begin{aligned} r_n(x_i) &= \frac{(x_n - x_i)p_{n-1}(x_i) + (x_i - x_0)q_{n-1}(x_i)}{(x_n - x_0)} \\ &= \frac{(x_n - x_i)f(x_i) + (x_i - x_0)f(x_i)}{(x_n - x_0)} = f(x_i) \end{aligned}$$

Les polynômes $p_n(x)$ et $r_n(x)$ passent par les mêmes $(n + 1)$ points; ils sont donc égaux, ce qui termine la démonstration du lemme.
L'équation 5.18 peut également s'écrire:

$$p_n(x) = \frac{(x_0 - x + x_n - x_0)p_{n-1}(x) + (x - x_0)q_{n-1}(x)}{(x_n - x_0)}$$

ou encore

$$p_n(x) - p_{n-1}(x) = \frac{(x - x_0)}{(x_n - x_0)}(q_{n-1}(x) - p_{n-1}(x)) \tag{5.19}$$

La démonstration du théorème vient ensuite assez facilement. Les coefficients de la puissance x^{n-1} pour les polynômes $p_{n-1}(x)$ et $q_{n-1}(x)$ sont respectivement a_{n-1} et b_{n-1} en vertu des équations 5.16 et 5.17. Selon l'équation 5.19, le coefficient de la puissance x^n de $p_n(x)$ est:

$$\frac{b_{n-1} - a_{n-1}}{x_n - x_0}$$

qui est a_n en vertu de la relation 5.15. La formule 5.19 permet aussi de trouver les racines de $p_n(x) - p_{n-1}(x)$, qui sont $x_0, x_1, \cdots, x_{n-1}$. Ce polynôme s'écrit donc:

$$p_n(x) - p_{n-1}(x) = a_n(x - x_0)(x - x_1) \cdots (x - x_{n-1})$$

ce qui termine la démonstration du théorème. □

Remarque 5.6

Une fois les coefficients a_i connus, on peut évaluer le polynôme de Newton au moyen d'un algorithme similaire au schéma d'Horner (voir la section 1.5.3). On écrit alors le polynôme 5.6 sous la forme:

$$\begin{aligned} p_n(x) &= a_0 + (x - x_0)(a_1 + (x - x_1)(a_2 + (x - x_2)(a_3 + \cdots \\ &\quad +(x - x_{n-2})(a_{n-1} + a_n(x - x_{n-1}) \cdots))) \end{aligned} \tag{5.20}$$

De cette façon, on réduit le nombre d'opérations nécessaires à l'évaluation du polynôme. De plus, cette forme est moins sensible aux effets des erreurs d'arrondis. □

Il reste maintenant à calculer efficacement la valeur de ce polynôme. La manière la plus simple consiste à construire une table dite de différences divisées de la façon suivante.

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$
x_0	$f(x_0)$			
		$f[x_0, x_1]$		
x_1	$f(x_1)$		$f[x_0, x_1, x_2]$	
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$
x_2	$f(x_2)$		$f[x_1, x_2, x_3]$	
		$f[x_2, x_3]$		
x_3	$f(x_3)$			

La construction de cette table est simple. Nous nous sommes arrêtés aux troisièmes différences divisées, mais les autres s'obtiendraient de la même manière. Les premières différences divisées découlent de la définition 5.8. Pour obtenir par exemple $f[x_0, x_1, x_2]$, il suffit de soustraire les 2 termes adjacents $f[x_1, x_2] - f[x_0, x_1]$ et de diviser le résultat par $(x_2 - x_0)$. De même, pour obtenir $f[x_0, x_1, x_2, x_3]$, on soustrait $f[x_0, x_1, x_2]$ de $f[x_1, x_2, x_3]$ et on divise le résultat par $(x_3 - x_0)$. La formule de Newton utilise la diagonale principale de cette table.

Exemple 5.5

La table de différences divisées pour les points $(0\,,\ 1), (1\,,\ 2), (2\,,\ 9)$ et $(3\,,\ 28)$ est:

x_i	$f(x_i)$	$f[x_i, , x_{i+1}]$	$f[x_i, \cdots, x_{i+2}]$	$f[x_i, \cdots, x_{i+3}]$
0	1			
		1		
1	2		3	
		7		1
2	9		6	
		19		
3	28			

Suivant la formule de Newton 5.6, avec $x_0 = 0$, le polynôme de collocation est:

$$p_3(x) = 1 + 1(x-0) + 3(x-0)(x-1) + 1(x-0)(x-1)(x-2) = x^3 + 1$$

qui est le même polynôme (en vertu de l'unicité) que celui obtenu par la méthode de Lagrange. On remarque de plus que le polynôme:

$$p_2(x) = 1 + 1(x-0) + 3(x-0)(x-1)$$

passe quant à lui par les trois premiers points de collocation. Si on souhaite ajouter un point de collocation et calculer un polynôme de degré 4, il n'est pas nécessaire de tout recommencer. Par exemple, si on veut inclure le point $(5, 54)$, on peut compléter la table de différences divisées déjà utilisée.

x_i	$f(x_i)$	$f[x_i, , x_{i+1}]$	$f[x_i, \cdots, x_{i+2}]$	$f[x_i, \cdots, x_{i+3}]$	$f[x_i, \cdots, x_{i+4}]$
0	1				
		1			
1	2		3		
		7		1	
2	9		6		$-3/5$
		19		-2	
3	28		-2		
		13			
5	54				

Ce polynôme de degré 4 est alors:

$$p_4(x) = p_3(x) + (-3/5)(x - 0)(x - 1)(x - 2)(x - 3)$$

qui est tout simplement le polynôme de degré 3 déjà calculé auquel on a ajouté une correction de degré 4.

• • • • •

Exemple 5.6

Il est bon de remarquer que les points de collocation ne doivent pas forcément être placés par abscisses croissantes, bien que cela soit souvent préférable. Considérons par exemple la table suivante:

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, \cdots, x_{i+2}]$	$f[x_i, \cdots, x_{i+3}]$
2	1			
		1		
0	−1		0,4	
		2,2		1,2
5	10		1,6	
		7		
3	−4			

On note que les abscisses x_i ne sont pas par ordre croissant. Le polynôme passant par ces points est:

$$p_3(x) = 1 + 1(x - 2) + 0{,}4(x - 2)(x - 0) + 1{,}2(x - 2)(x - 0)(x - 5)$$

que l'on obtient de la relation 5.6 en prenant $x_0 = 2$. Si on souhaite évaluer ce polynôme en $x = 1$, on peut se servir de la méthode de Horner. On réécrit alors le polynôme sous la forme:

$$p_3(x) = 1 + (x - 2)(1 + (x - 0)(0{,}4 + 1{,}2(x - 5)))$$

La fonction inconnue $f(x)$ peut alors être estimée par ce polynôme. Ainsi:

$$\begin{aligned} f(1) \simeq p_3(1) &= 1 + (-1)(1 + (1)(0{,}4 + 1{,}2(-4))) \\ &= 1 + (-1)(1 - 4{,}4) \\ &= 4{,}4 \end{aligned}$$

Pour l'instant, nous n'avons aucune indication quant à la précision de cette approximation. Cette question est l'objet de la section suivante.

●●●●●

5.5 Erreur d'interpolation

L'interpolation permet, à partir d'un certain nombre de données sur les valeurs d'une fonction, de faire l'approximation de $f(x)$ en tout point x.

Toutefois, cette opération entraîne une *erreur d'interpolation* qu'il convient d'étudier en détail, d'autant plus que les résultats nous serviront également dans l'analyse de l'intégration et de la dérivation numériques.

On peut exprimer l'erreur d'interpolation de la façon suivante:

$$f(x) = p_n(x) + E_n(x)$$

ou encore

$$E_n(x) = f(x) - p_n(x)$$

Cela signifie que le polynôme $p_n(x)$ de degré n procure une approximation de la fonction $f(x)$ avec une erreur $E_n(x)$. Il reste à évaluer cette erreur. On constate immédiatement que:

$$E_n(x_i) = 0 \quad \text{pour } i = 0, 1, 2, \cdots, n$$

et donc que l'erreur d'interpolation est nulle aux points de collocation puisque le polynôme passe exactement par ces points.

Remarque 5.7

On suppose que les données des points $(x_i, f(x_i))$ sont exactes, ce qui n'est pas toujours le cas. En effet, si ces données proviennent de mesures expérimentales, elles peuvent être entachées d'une erreur de mesure. Dans ce qui suit, nous supposons que cette erreur est nulle. □

Le résultat suivant donne une expression analytique du terme d'erreur (voir Fortin et Pierre, réf. [11], pour une démonstration).

Théorème 5.4

Soit $x_0 < x_1 < x_2 < \cdots < x_n$, des points de collocation. On suppose que la fonction $f(x)$ est définie dans l'intervalle $[x_0, x_n]$ et qu'elle est $(n+1)$ fois dérivable dans (x_0, x_n). Alors, pour tout x compris dans $[x_0, x_n]$, il existe $\xi(x)$ appartenant à l'intervalle (x_0, x_n) tel que:

$$E_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n) \tag{5.21}$$

La relation 5.21 est l'*expression analytique de l'erreur d'interpolation.* □

Plusieurs commentaires sont nécessaires pour bien comprendre ce résultat.

- On constate immédiatement que $E_n(x_i) = 0$ quel que soit i choisi entre 0 et n. L'erreur d'interpolation est nulle aux points de collocation.

- La fonction *a priori* inconnue $f(x)$ apparaît par l'entremise de sa dérivée d'ordre $(n+1)$ évaluée au point $\xi(x)$, également inconnu et qui varie avec x.

- Il existe une similarité entre l'erreur d'interpolation et l'erreur reliée au développement de Taylor 1.25. Dans les deux cas, on montre l'existence d'un point $\xi(x)$ permettant d'évaluer l'erreur mais que l'on ne peut généralement pas déterminer.

- *Puisque le terme d'erreur en un point x fait intervenir des coefficients de la forme $(x - x_i)$, il y a tout intérêt à choisir les points x_i qui sont situés le plus près possible de x. Ce choix est utile lorsqu'un grand nombre de points de collocation sont disponibles et qu'il n'est pas nécessaire de construire un polynôme passant par tous les points. On retient alors seulement les points de collocation les plus près de x de manière à minimiser l'erreur.*

- La fonction $(x-x_0)(x-x_1)\cdots(x-x_n)$ est un polynôme de degré $(n+1)$ et possède donc les $(n+1)$ racines réelles (x_i pour $i = 0, 1, \cdots, n$). Dans certaines conditions, cette fonction peut osciller avec de fortes amplitudes, d'où le risque de grandes erreurs d'interpolation. *Cette propriété fait en sorte qu'il est délicat d'effectuer des interpolations en utilisant des polynômes de degré élevé.*

Exemple 5.7
Soit les valeurs expérimentales suivantes, que l'on a obtenues en mesurant la vitesse (en km/h) d'un véhicule toutes les 5 secondes:

$$\begin{array}{ccccc} (0\,,\,55) & (5\,,\,60) & (10\,,\,58) & (15\,,\,54) & (20\,,\,55) \\ (25\,,\,60) & (30\,,\,54) & (35\,,\,57) & (40\,,\,52) & (45\,,\,49) \end{array}$$

On constate que le véhicule se déplace à une vitesse oscillant autour de 55 km/h. On peut établir l'équation d'un polynôme de degré 9 passant par ces dix points (voir la figure 5.4). On remarque de fortes oscillations de ce polynôme principalement au début et à la fin de l'intervalle $[0, 45]$. Ainsi, si

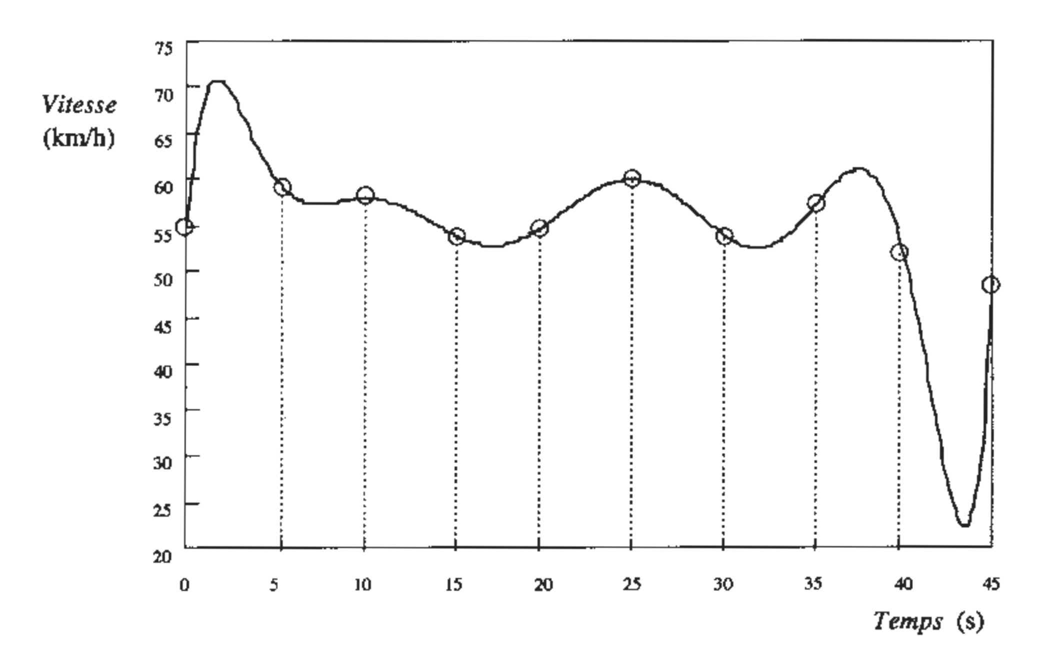


Figure 5.4: Interpolation de la vitesse d'un véhicule

on interpole les valeurs en $t = 2{,}5$ s et en $t = 42{,}5$ s, on trouve des vitesses respectives de 69,248 km/h et de 27,02 km/h, ce qui semble peu probable puisque le véhicule se déplace à une vitesse à peu près uniforme. Si on regarde les valeurs adjacentes au temps $t = 42{,}5$ s, le véhicule serait passé de 52 km/h à 49 km/h en passant par un creux soudain de 27,02 km/h. Rien ne laisse supposer un tel comportement. Pour remédier à la situation, on doit recourir à des polynômes de degré plus faible. Ainsi, en prenant seulement les trois premiers points de collocation qui définissent un polynôme de degré 2, on trouve une vitesse de 58,375 km/h en $t = 2{,}5$ s. De même, en ne considérant que les trois derniers points de collocation, on trouve une vitesse de 50,25 km/h en $t = 42{,}5$ s. Ces résultats sont beaucoup plus acceptables.

Profitons de l'occasion pour souligner les dangers de l'*extrapolation*, c'est-à-dire de l'utilisation du polynôme d'interpolation à l'extérieur de l'intervalle contenant les points de collocation. Dans cet exemple où l'intervalle est [0, 45], on trouverait une vitesse de 256 km/h en $t = 47$ s et de 1635 km/h en $t = 50$ s. Ces valeurs sont inutilisables.

● ● ● ● ●

Remarque 5.8
L'expression analytique du terme d'erreur nous impose de choisir les points d'interpolation les plus près du point x où l'on veut interpoler. Cela s'est avéré souhaitable dans l'exemple précédent et est en fait une règle générale. □

Exemple 5.8
Soit les points (1 , 1), (3 , 1,732 051), (7,5 , 2,738 613), (9,1 , 3,016 620) et (12 , 3,464 102). Si on veut interpoler la fonction inconnue $f(x)$ en $x = 8$, il est utile de construire une table de différences divisées.

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, \cdots, x_{i+2}]$	$f[x_i, \cdots, x_{i+3}]$	$f[x_i, \cdots, x_{i+4}]$
7,5	2,738 613				
		0,173 755			
9,1	3,016 621		−0,004 322 47		
		0,154 304		0,000 4291	
12	3,464 102		−0,006 253 44		0,000 1149
		0,192 450		0,001 1761	
3	1,732 051		−0,015 779 54		
		0,366 025			
1	1,0				

On remarque que les abscisses x_i ont été ordonnées en fonction de leur distance par rapport à $x = 8$. Cela permet d'effectuer d'abord l'interpolation avec les valeurs les plus proches de x et également de diminuer plus rapidement l'erreur d'interpolation. En effet, en prenant des polynômes de degré de plus en plus élevé, la formule de Newton donne les résultats suivants.

Degré n	$p_n(8)$	$\lvert p_n(8) - \sqrt{8}\rvert$
1	2,825 490	$0{,}29 \times 10^{-2}$
2	2,827 868	$0{,}55 \times 10^{-3}$
3	2,828 812	$0{,}38 \times 10^{-3}$
4	2,827 547	$0{,}88 \times 10^{-3}$

La fonction interpolée est $f(x) = \sqrt{x}$, qui prend la valeur de 2,828 427 125 en $x = 8$. On constate donc une précision acceptable dès le polynôme de degré 1. Si les points d'interpolation avaient été classés par abscisse croissante, on aurait obtenu le tableau suivant.

Degré n	$p_n(8)$	$\lvert p_n(8) - \sqrt{8}\rvert$
1	3,562 178	$0,73 \times 10^{0}$
2	2,795 705	$0,32 \times 10^{-1}$
3	2,825 335	$0,30 \times 10^{-2}$
4	2,827 547	$0,88 \times 10^{-3}$

Il faut dans ce cas attendre le degré 3 avant d'avoir une précision acceptable. On obtient bien sûr le même résultat dans les deux cas lorsque tous les points d'interpolation sont utilisés, c'est-à-dire lorsqu'on recourt au polynôme de degré 4. On voit donc l'importance d'utiliser les points de collocation les plus près possible de l'abscisse autour de laquelle on veut effectuer l'interpolation.

• • • • •

L'expression analytique de l'erreur d'interpolation 5.21 ne permet pas d'évaluer la précision de l'approximation. Il est cependant souhaitable de pouvoir évaluer cette erreur, même de façon grossière. Cela est possible avec la formule de Newton. En effet, l'expression 5.21 fait intervenir la dérivée d'ordre $(n+1)$ de la fonction $f(x)$ en $x = \xi$. C'est ce terme qu'il est nécessaire d'estimer, puisque c'est le seul qui ne puisse être évalué exactement.

Considérons le cas particulier où les abscisses x_i sont également distantes, c'est-à-dire où:

$$x_{i+1} - x_i = h$$

Il faut établir un lien entre les dérivées de la fonction $f(x)$ et les différences divisées. On remarque dans un premier temps que $f[x_0, x_1]$ est une approximation d'ordre 1 de la dérivée de $f(x)$ en $x = x_0$:

$$f[x_0, x_1] = f'(x_0) + O(h)$$

En effet, on a:

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{(x_1 - x_0)} = \frac{f(x_0 + h) - f(x_0)}{h}$$

En utilisant le développement de Taylor 1.24, on obtient:

$$\begin{aligned} f[x_0, x_1] &= \frac{(f(x_0) + f'(x_0)h + f''(x_0)h^2/2 + O(h^3)) - f(x_0)}{h} \\ &= f'(x_0) + \frac{f''(x_0)h}{2} + O(h^2) = f'(x_0) + O(h) \end{aligned}$$

De même, on peut montrer qu'à une constante près la n^e différence divisée de $f(x)$ est une approximation d'ordre 1 de la dérivée n^e de $f(x)$ en $x = x_0$. On peut en effet démontrer qne:

$$f[x_0, x_1, x_2 \cdots, x_n] = \frac{f^{(n)}(x_0)}{n!} + O(h) \tag{5.22}$$

Nous reviendrons sur l'approximation 5.22 au chapitre 6 sur la dérivation numérique. Pour le moment, servons-nous de ce résultat. On suppose que la dérivée $(n+1)^e$ de $f(x)$ varie pen dans l'intervalle $[x_0, x_n]$. On a alors l'approximation suivante:

$$f[x_0, x_1, x_2, \cdots, x_n, x_{n+1}] \simeq \frac{f^{(n+1)}(x_0)}{(n+1)!} \simeq \frac{f^{n+1}(\xi)}{(n+1)!}$$

On peut ainsi estimer le terme d'erreur 5.21 par:

$$E_n(x) \simeq f[x_0, x_1, x_2, \cdots, x_n, x_{n+1}](x - x_0)(x - x_1)\cdots(x - x_n) \tag{5.23}$$

Remarque 5.9
On remarque immédiatement que l'approximation 5.23 n'est rien d'autre que le terme nécessaire au calcul du polynôme de degré $(n + 1)$ dans la formule de Newton 5.6. En d'autres termes, il est possible d'évaluer l'erreur d'interpolation liée à un polynôme de degré n en calculant le terme suivant dans la formule de Newton.

L'approximation 5.23 n'est pas toujours d'une grande précision, mais c'est généralement la seule disponible. □

Cela nous amène à suggérer le critère d'arrêt suivant dans le cas de l'interpolation à l'aide de la formule de Newton. On considère que l'approximation $p_n(x)$ est suffisamment précise si:

$$\frac{|p_{n+1}(x) - p_n(x)|}{|p_{n+1}(x)|} < \epsilon$$

où ϵ est une valeur de tolérance fixée à l'avance. Il est généralement recommandé de fixer également le degré maximal N des polynômes utilisés.

Exemple 5.9
Soit une table de la fonction $\sqrt{x}$. Puisqu'on connaît la fonction (ce qui n'est bien sûr pas le cas en pratique), on est donc en mesure d'évaluer l'erreur exacte et de la comparer avec son approximation obtenue à l'aide de la relation 5.23.

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, \cdots, x_{i+2}]$	$f[x_i, \cdots, x_{i+3}]$	$f[x_i, \cdots, x_{i+4}]$
7	2,645 751				
		0,177 124			
9	3,000 000		−0,004 702 99		
		0,158 312		0,000 206 783	
11	3,316 625		−0,003 462 29		$0{,}9692 \times 10^{-5}$
		0,144 463		0,000 129 248	
13	3,605 551		−0,002 686 80		
		0,133 716			
15	3,872 983				

On tente d'obtenir une approximation de $\sqrt{8}$ à l'aide de cette table. En se basant sur un polynôme de degré 1 et en prenant $x_0 = 7$, on obtient facilement:

$$p_1(x) = 2{,}645\,751 + 0{,}177\,124(x - 7)$$

de telle sorte que:

$$p_1(8) = 2{,}822\,875$$

L'erreur exacte en $x = 8$ est alors:

$$E_1(8) = f(8) - p_1(8) = \sqrt{8} - 2{,}822\,875 = 0{,}005\,552\,125$$

Selon l'expression 5.23, on peut estimer cette erreur par le terme suivant dans la formule de Newton 5.6, c'est-à-dire:

$$E_1(8) \simeq -0{,}004\,702\,99(8-7)(8-9) = 0{,}004\,702\,99$$

On constate donc que l'erreur approximative est assez près de l'erreur exacte. Considérons maintenant le polynôme de degré 2:

$$p_2(x) = p_1(x) - 0{,}004\,702\,99(x-7)(x-9)$$

qui prend la valeur:

$$p_2(8) = 2{,}822\,875 + 0{,}004\,702\,99 = 2{,}827\,577\,990$$

soit une erreur exacte de 0,000 849 135. Encore ici, cette erreur peut être approchée à l'aide du terme suivant dans la formule de Newton:

$$E_2(8) \simeq 0{,}000\,206\,783(8-7)(8-9)(8-11) = 0{,}000\,620\,349$$

Enfin, en passant au polynôme de degré 3, on trouve:

$$p_3(x) = p_2(x) + 0{,}000\,206\,783(x-7)(x-9)(x-11)$$

ce qui entraîne que:

$$p_3(8) = 2{,}827\,578\,301 + 0{,}000\,620\,349 = 2{,}828\,198\,339$$

L'erreur exacte est alors 0,000 228 786, ce qui est près de la valeur obtenue au moyen de l'équation 5.23:

$$E_3(8) \simeq 0{,}9692 \times 10^{-5}(8-7)(8-9)(8-11)(8-13) = 0{,}000\,145\,380$$

qui montre que cette approximation possède 4 chiffres significatifs.

On remarque par ailleurs dans la table que les premières différences divisées sont négatives et que le signe alterne d'une colonne à une autre. Cela s'explique par la relatiou 5.22, qui établit un lien entre les différences divisées et les dérivées de la fonction $f(x)$. Dans cet exemple, on a:

$$f(x) = \sqrt{x}, \quad f'(x) = \frac{1}{2\sqrt{x}}, \quad f''(x) = \frac{-1}{4x^{3/2}}, \quad f'''(x) = \frac{3}{8x^{5/2}}, \quad \text{etc.}$$

Le signe des dérivées alterne, tout comme le signe des différentes colonnes de la table de différences divisées.

● ● ● ● ●

Nous terminerons cette section en cherchant à déterminer l'ordre de convergence de l'approximation polynomiale. Si on retient le cas où les abscisses sont également distantes, il suffit de poser:

$$s = \frac{x - x_0}{h} \quad \text{ou encore } (x - x_0) = sh \tag{5.24}$$

On remarque alors que:

$$x - x_i = x - (x_0 + ih) = (x - x_0) - ih = sh - ih = (s - i)h$$

Il suffit maintenant de remplacer $x - x_i$ par $(s - i)h$ dans l'expression analytique de l'erreur d'interpolation 5.21 pour obtenir le prochain résultat.

Théorème 5.5

Dans le cas où les points de collocation x_i sont équidistants, l'expression analytique de l'erreur d'interpolation s'écrit:

$$E_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} s(s-1)(s-2)\cdots(s-n)h^{n+1} \qquad (5.25)$$

pour un certain ξ dans l'intervalle (x_0, x_n) et pour s défini par l'équation 5.24.
□

Remarque 5.10

On peut dès lors conclure que le polynôme d'interpolation $p_n(x)$ est une approximation d'ordre $(n + 1)$ de la fonction $f(x)$. Encore une fois, si on prend des points de collocation situés à une distance $h/2$ les uns des autres, l'erreur d'interpolation est diminuée d'un facteur de 2^{n+1}. □

Exemple 5.10

Pour illustrer l'ordre de convergence d'un polynôme de collocation, il est nécessaire de regarder le comportement de l'erreur lorsque h tend vers 0. Par exemple, on peut utiliser un polynôme de degré 2 (précis à l'ordre 3) construit à partir des points:

$$(x_0 - h/2, f(x_0 - h/2))$$
$$(x_0 + h/2, f(x_0 + h/2))$$
$$(x_0 + 3h/2, f(x_0 + 3h/2))$$

faire tendre h vers 0 et vérifier la décroissance de l'erreur en $x = x_0$. Le problème est illustré à la figure 5.5, où on constate que le point x_0 est coincé entre les 3 abscisses choisies, quel que soit h.

On a obtenu le tableau qui suit en utilisant la fonction $f(x) = e^x$ au point $x_0 = 2{,}6$.

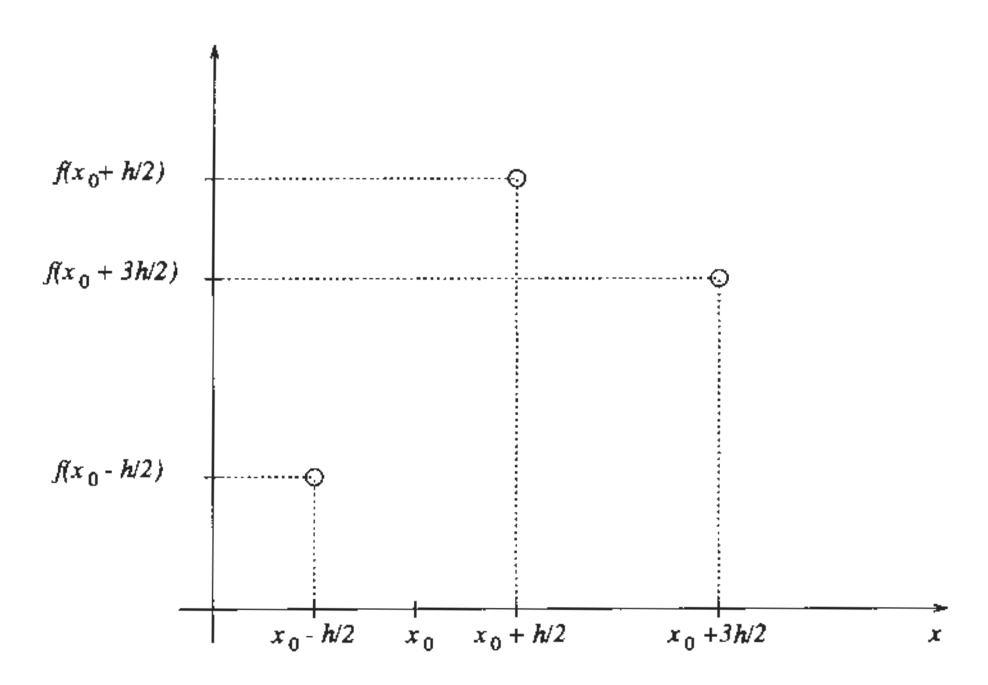


Figure 5.5: Trois points de collocation

h	$p_2(2{,}6)$	$\|p_2(2{,}6) - e^{2{,}6}\|$	$e(2h)/e(h)$
$0{,}500\,0000 \times 10^{0}$	13,335 083	$0{,}128\,65 \times 10^{0}$	—
$0{,}250\,0000 \times 10^{0}$	13,449 248	$0{,}144\,90 \times 10^{-1}$	8, 878 61
$0{,}125\,0000 \times 10^{0}$	13,462 014	$0{,}172\,38 \times 10^{-2}$	8, 405 61
$0{,}625\,0000 \times 10^{-1}$	13,463 528	$0{,}210\,35 \times 10^{-3}$	8, 194 91
$0{,}312\,5000 \times 10^{-1}$	13,463 712	$0{,}259\,84 \times 10^{-4}$	8, 095 61
$0{,}156\,2500 \times 10^{-1}$	13,463 735	$0{,}322\,89 \times 10^{-5}$	8, 047 31
$0{,}781\,2500 \times 10^{-2}$	13,463 738	$0{,}402\,42 \times 10^{-6}$	8, 023 51
$0{,}390\,6250 \times 10^{-2}$	13,463 738	$0{,}502\,29 \times 10^{-7}$	8, 011 71
$0{,}195\,3125 \times 10^{-2}$	13,463 738	$0{,}627\,41 \times 10^{-8}$	8, 005 81
$0{,}976\,5625 \times 10^{-3}$	13,463 738	$0{,}783\,97 \times 10^{-9}$	8, 002 91

On remarque que la valeur de h est systématiquement divisée par 2 et que l'erreur tend vers 0. De plus, le ratio de l'erreur liée à la valeur de h précédente (deux fois plus grande) et de l'erreur liée à la valeur de h actuelle fait apparaître la valeur de 8 (ou 2^3), ce qui confirme l'ordre 3 de cette approximation.

● ● ● ● ●

Remarque 5.11

- L'expression analytique de l'erreur d'interpolation demeure la même, quelle que soit la façon dont on calcule le polynôme d'interpolation. Ainsi, l'expression 5.21 est valable si on utilise l'interpolation de Lagrange, la matrice de Vandermonde ou toute autre méthode. Cela s'explique par l'unicité de ce polynôme.

- L'approximation de l'erreur exprimée par l'équation 5.23 est également valable quelle que soit la façon dont on calcule le polynôme d'interpolation. Cependant, si on utilise une autre méthode que l'interpolation de Newton, il faut calculer la table de différences divisées pour obtenir l'approximation 5.23. Il est donc avantageux d'utiliser la formule de Newton dès le départ. □

5.6 Splines cubiques

Nous avons constaté que l'utilisation de polynômes de degré élevé est parfois délicate et peut mener à des erreurs d'interpolation importantes. De plus, il est parfois nécessaire d'obtenir des courbes très régulières passant par un grand nombre de points. C'est le cas en conception assistée par ordinateur (CAO), où l'on cherche à représenter des objets aux formes régulières. Les polynômes de degré élevé sont alors peu adéquats.

On peut mesurer la régularité d'une fonction par le biais de ses dérivées. En effet, plus une fonction est différentiable, plus la courbe qui lui est associée est lisse et plus la fonction est régulière. Le problème, lorsqu'on utilise des polynômes de faible degré, provient du fait qu'il faut en utiliser plusieurs pour relier tous les points. C'est le cas de l'interpolation linéaire par morceaux, illustrée à la figure 5.6, qui consiste à relier chaque paire de points par un segment de droite. On utilise aussi l'appellation *splines linéaires*. On imagine assez mal comment une telle courbe pourrait permettre de faire le design d'une carrosserie de voiture ou d'une aile d'avion.

Il faut donc être plus prudent à la jonction de ces différents segments de courbe. Une voie très populaire consiste à utiliser dans chaque intervalle $[x_{i-1}, x_i]$ un polynôme de degré 3 de la forme:

$$p_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i \quad \text{pour } i = 1, 2, \cdots, n$$

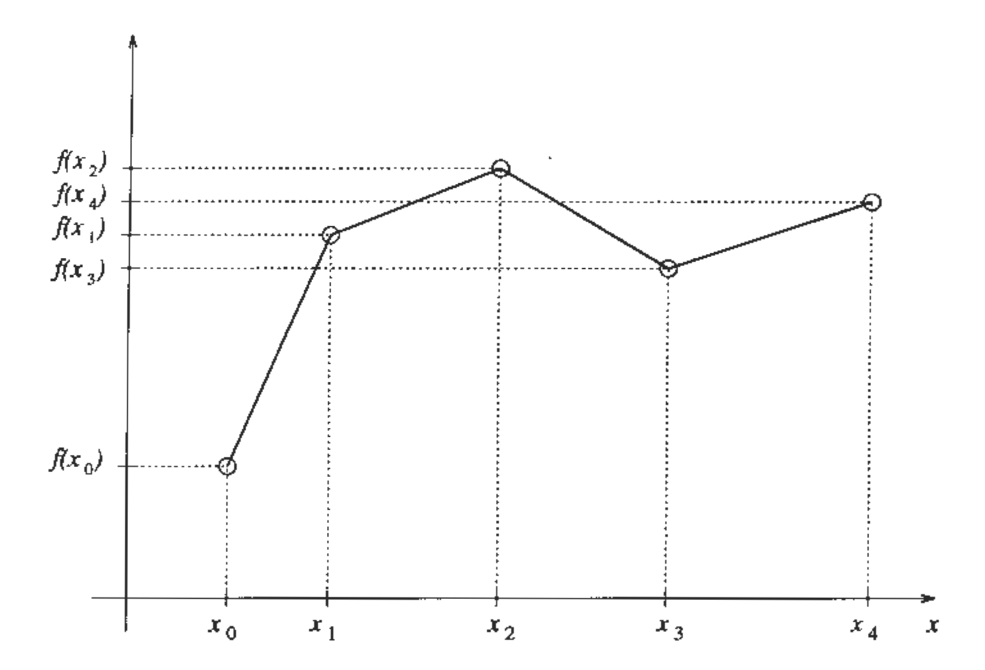


Figure 5.6: Interpolation linéaire par morceaux

et à relier ces différents polynômes de façon à ce que la courbe résultante soit deux fois différentiable. C'est l'*interpolation par splines cubiques.* Supposons que l'on a $(n+1)$ points d'interpolation et donc n intervalles $[x_{i-1}, x_i]$. Cela indique qu'il y a $4n$ coefficients (a_i, b_i, c_i, d_i pour $i = 1, 2, \cdots, n$) à déterminer. La situation est décrite à la figure 5.7 pour $n = 4$.

Ces $4n$ coefficients doivent être déterminés le plus efficacement possible pour que la méthode reste attrayante. Une résolution astucieuse conduit à un système linéaire tridiagonal de dimension $(n-1)$. On sait que les systèmes tridiagonaux sont très faciles à résoudre puisque la décomposition LU se réduit dans ce cas à peu d'opérations (voir Chapra et Canale, réf. [4]).

Voyons combien de conditions ou d'équations nous pouvons imposer à ces $4n$ coefficients. Ces équations proviennent des conditions de régularité que l'on souhaite imposer à la courbe résultante. Pour faciliter la compréhension, il est préférable de distinguer parmi les points d'interpolation les deux extrémités et les $(n-1)$ points intérieurs. Les deux extrémités sont les points $(x_0, f(x_0))$ et $(x_n, f(x_n))$. Une attention particulière doit être portée aux points intérieurs, qui se trouvent à la jonction de deux polynômes de degré 3. Voici les contraintes imposées aux n polynômes de degré 3:

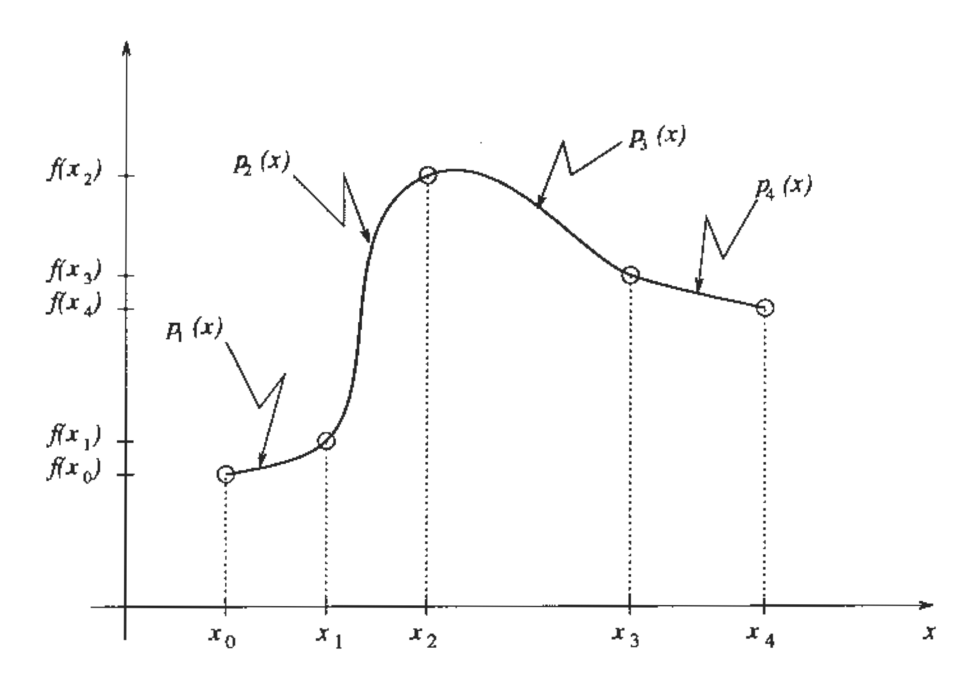


Figure 5.7: Splines cubiques: n polynômes de degré 3

- Le polynôme $p_1(x)$ passe par la première extrémité $(x_0, f(x_0))$, c'est-à-dire:

$$p_1(x_0) = f(x_0)$$

  et de même à l'autre extrémité:

$$p_n(x_n) = f(x_n)$$

  ce qui introduit 2 équations.

- Par chaque point intérieur (x_i pour $i = 1, 2, \cdots, n-1$) passent deux polynômes, soit $p_i(x)$ défini dans l'intervalle $[x_{i-1}, x_i]$ et $p_{i+1}(x)$ défini dans $[x_i, x_{i+1}]$. Ces deux polynômes doivent passer par le point $(x_i, f(x_i))$, c'est-à-dire:

$$p_i(x_i) = f(x_i) \quad \text{pour } i = 1, 2, \cdots, n-1$$

  et

$$p_{i+1}(x_i) = f(x_i) \quad \text{pour } i = 1, 2, \cdots, n-1$$

  Cela résulte en $2(n-1) = (2n-2)$ équations supplémentaires.

- Pour assurer la régularité de la courbe, on doit imposer que les dérivées premières et secondes de ces deux polynômes soient continues aux points intérieurs. On doit donc imposer:

$$p_i'(x_i) = p_{i+1}'(x_i) \text{ pour } i = 1, 2, \cdots, n-1 \tag{5.26}$$

  et

$$p_i''(x_i) = p_{i+1}''(x_i) = f_i'' \text{ pour } i = 1, 2, \cdots, n-1 \tag{5.27}$$

  ce qui donne $(2n-2)$ nouvelles équations. Nous avons introduit la notation f_i'' pour désigner la valeur de la dérivée seconde de la spline au point intérieur x_i. De même, nous désignons f_0'' et f_n'' les valeurs de la dérivée seconde aux deux extrémités.

Au total, on a $(4n-2)$ équations en $4n$ inconnues et il manque donc 2 équations pour pouvoir résoudre ce système linéaire. Comme nous le verrons plus loin, il existe plusieurs façons de rajouter ces 2 équations.

Voyons maintenant comment déterminer l'équation de la spline $p_i(x)$ dans chacun des intervalles $[x_{i-1}, x_i]$. Il est fort heureusement possible de ramener ce système de $(4n-2)$ équations en $4n$ inconnues en un système beaucoup plus petit de $(n-1)$ équations en $(n-1)$ inconnues. Ces $(n-1)$ inconnues sont tout simplement les valeurs des dérivées secondes (f_i'') de la spline aux points d'interpolation.

La résolution du système est basée sur la constatation suivante: puisque la spline est constituée de polynômes de degré 3 dans chaque intervalle, la dérivée seconde de la spline est un polynôme de degré 1 dans chaque intervalle. De plus, cette dérivée seconde étant continue en vertu de l'équation 5.27, on a la situation de la figure 5.8 où la dérivée seconde est représentée par des segments de droite dont les sommets sont les points (x_i, f_i'').

Ainsi, dans l'intervalle $[x_{i-1}, x_i]$, la dérivée seconde $p_i''(x)$ est un polynôme de degré 1 dont les sommets sont (x_{i-1}, f_{i-1}'') et (x_i, f_i''). La formule d'interpolation de Lagrange donne alors:

$$p_i''(x) = f_{i-1}'' \frac{(x - x_i)}{(x_{i-1} - x_i)} + f_i'' \frac{(x - x_{i-1})}{(x_i - x_{i-1})}$$

En intégrant deux fois cette équation, on obtient d'abord:

$$p_i'(x) = f_{i-1}'' \frac{(x - x_i)^2}{2(x_{i-1} - x_i)} + f_i'' \frac{(x - x_{i-1})^2}{2(x_i - x_{i-1})} + \alpha_i$$

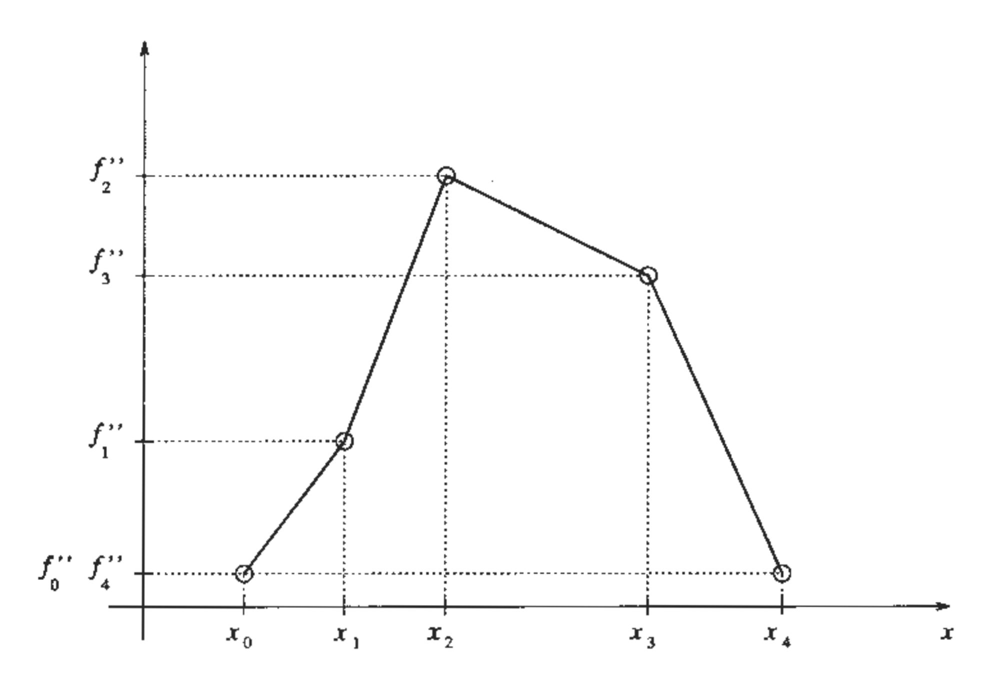


Figure 5.8: Dérivée seconde de la spline

et par la suite:

$$p_i(x) = f''_{i-1}\frac{(x-x_i)^3}{6(x_{i-1}-x_i)} + f''_i\frac{(x-x_{i-1})^3}{6(x_i-x_{i-1})} + \alpha_i(x-x_i) + \beta_i$$

où les α_i et les β_i sont des constantes d'intégration. Dans une première étape, on exprime ces constantes d'intégration en fonction des inconnues f''_i.

La courbe du polynôme $p_i(x)$ défini dans $[x_{i-1}, x_i]$ doit passer par les points $(x_{i-1}, f(x_{i-1}))$ et $(x_i, f(x_i))$. On en déduit que:

$$p_i(x_i) = f(x_i) = f''_i\frac{(x_i-x_{i-1})^3}{6(x_i-x_{i-1})} + \beta_i = f''_i\frac{(x_i-x_{i-1})^2}{6} + \beta_i$$

ce qui entraîne que:

$$\beta_i = f(x_i) - f''_i\frac{(x_i-x_{i-1})^2}{6} \tag{5.28}$$

De même, en $x = x_{i-1}$:

$$p_i(x_{i-1}) = f(x_{i-1}) = f''_{i-1}\frac{(x_{i-1}-x_i)^3}{6(x_{i-1}-x_i)} + \alpha_i(x_{i-1}-x_i) + \beta_i$$

d'où l'on tire, d'après la relation 5.28:

$$f(x_{i-1}) = f''_{i-1}\frac{(x_{i-1}-x_i)^2}{6} + \alpha_i(x_{i-1}-x_i) + f(x_i) - f''_i\frac{(x_i-x_{i-1})^2}{6}$$

ce qui permet d'isoler α_i pour obtenir:

$$\alpha_i = \frac{f(x_i)-f(x_{i-1})}{(x_i-x_{i-1})} - \frac{(f''_i - f''_{i-1})(x_i-x_{i-1})}{6} \tag{5.29}$$

On en déduit que l'équation de la spline dans l'intervalle $[x_{i-1}, x_i]$ est:

$$\begin{aligned} p_i(x) &= f''_{i-1}\frac{(x-x_i)^3}{6(x_{i-1}-x_i)} + f''_i\frac{(x-x_{i-1})^3}{6(x_i-x_{i-1})} \\ &+ \left(\frac{f(x_i)-f(x_{i-1})}{(x_i-x_{i-1})} - \frac{(f''_i-f''_{i-1})(x_i-x_{i-1})}{6}\right)(x-x_i) \\ &+ f(x_i) - f''_i\frac{(x_i-x_{i-1})^2}{6} \end{aligned}$$

ou encore

$$\begin{aligned} p_i(x) &= f''_{i-1}\frac{(x-x_i)^3}{6(x_{i-1}-x_i)} + f''_i\frac{(x-x_{i-1})^3}{6(x_i-x_{i-1})} \\ &- \left(\frac{f(x_{i-1})}{(x_i-x_{i-1})} - \frac{f''_{i-1}(x_i-x_{i-1})}{6}\right)(x-x_i) \\ &+ \left(\frac{f(x_i)}{(x_i-x_{i-1})} - \frac{f''_i(x_i-x_{i-1})}{6}\right)(x-x_i) \\ &+ f(x_i) - f''_i\frac{(x_i-x_{i-1})^2}{6} \end{aligned}$$

Un dernier effort est nécessaire pour obtenir une forme plus compacte. Il suffit de remplacer dans le quatrième terme à droite le monôme $(x - x_i)$ par

l'expression équivalente $(x - x_{i-1}) + (x_{i-1} - x_i)$. On obtient alors:

$$\begin{aligned} p_i(x) &= -f''_{i-1}\frac{(x - x_i)^3}{6(x_i - x_{i-1})} + f''_i\frac{(x - x_{i-1})^3}{6(x_i - x_{i-1})} \\ &- \left(\frac{f(x_{i-1})}{(x_i - x_{i-1})} - f''_{i-1}\frac{(x_i - x_{i-1})}{6}\right)(x - x_i) \\ &+ \left(\frac{f(x_i)}{(x_i - x_{i-1})} - f''_i\frac{(x_i - x_{i-1})}{6}\right)(x - x_{i-1}) \end{aligned}$$

On peut simplifier quelque peu l'équation précédente en posant:

$$h_i = x_i - x_{i-1} \quad \text{pour } i = 1, 2, \cdots, n$$

ce qui permet d'obtenir:

$$\begin{aligned} p_i(x) &= -f''_{i-1}\frac{(x - x_i)^3}{6h_i} + f''_i\frac{(x - x_{i-1})^3}{6h_i} \\ &- \left(\frac{f(x_{i-1})}{h_i} - \frac{h_i f''_{i-1}}{6}\right)(x - x_i) \\ &+ \left(\frac{f(x_i)}{h_i} - \frac{h_i f''_i}{6}\right)(x - x_{i-1}) \end{aligned} \tag{5.30}$$

qui est l'*équation de la spline dans l'intervalle* $[x_{i-1}, x_i]$.

Il reste à déterminer les f''_i. Des $(4n - 2)$ conditions retenues, seule la continuité de la première dérivée (voir l'équation 5.26) n'a pas encore été imposée. Pour ce faire, il faut dériver $p_i(x)$ dans l'intervalle $[x_{i-1}, x_i]$ et $p_{i+1}(x)$ dans $[x_i, x_{i+1}]$, puis évaluer $p_i(x_i)$ et $p_{i+1}(x_i)$. On a d'une part:

$$\begin{aligned} p'_i(x) &= -f''_{i-1}\frac{(x - x_i)^2}{2h_i} + f''_i\frac{(x - x_{i-1})^2}{2h_i} \\ &- \left(\frac{f(x_{i-1})}{h_i} - \frac{h_i f''_{i-1}}{6}\right) \\ &+ \left(\frac{f(x_i)}{h_i} - \frac{h_i f''_i}{6}\right) \end{aligned}$$

et d'autre part:

$$\begin{aligned} p'_{i+1}(x) &= -f''_i \frac{(x - x_{i+1})^2}{2h_{i+1}} + f''_{i+1} \frac{(x - x_i)^2}{2h_{i+1}} \\ &\quad - \left(\frac{f(x_i)}{h_{i+1}} - \frac{h_{i+1} f''_i}{6} \right) \\ &\quad + \left(\frac{f(x_{i+1})}{h_{i+1}} - \frac{h_{i+1} f''_{i+1}}{6} \right) \end{aligned}$$

En égalisant ces deux expressions évaluées en $x = x_i$ et en les simplifiant, on obtient les $(n-1)$ équations suivantes:

$$\begin{aligned} h_i f''_{i-1} + 2(h_i + h_{i+1}) f''_i + h_{i+1} f''_{i+1} &= 6\left(f[x_i, x_{i+1}] - f[x_{i-1}, x_i]\right) \\ &\quad \text{pour } i = 1, 2, \cdots, n-1 \end{aligned}$$

Une dernière simplification est possible si on divise chaque terme de cette dernière équation par:

$$h_i + h_{i+1} = x_i - x_{i-1} + x_{i+1} - x_i = x_{i+1} - x_{i-1}$$

ce qui donne:

$$\begin{aligned} \frac{h_i}{(h_i + h_{i+1})} f''_{i-1} + 2f''_i + \frac{h_{i+1}}{(h_i + h_{i+1})} f''_{i+1} &= 6f[x_{i-1}, x_i, x_{i+1}] \\ \text{pour } \; i = 1, 2, 3, \cdots, n-1 & \end{aligned} \tag{5.31}$$

On remarque que le terme de droite fait intervenir les deuxièmes différences divisées. Il y a au total $(n+1)$ inconnues f''_i et on n'a que $(n-1)$ équations. On doit donc fixer de façon arbitraire deux des inconnues. Il existe plusieurs possibilités, mais la plus simple consiste à imposer:

$$f''_0 = f''_n = 0$$

On qualifie de *spline naturelle* la courbe qui en résulte. La spline naturelle impose que la dérivée seconde est nulle aux deux extrémités et donc que la courbe y est linéaire. Un autre choix possible consiste à imposer que:

$$f''_0 = f''_1 \quad \text{et} \quad f''_n = f''_{n-1}$$

ce qui revient à imposer une courbure constante dans le premier et dans le dernier intervalle.

Remarque 5.12

D'autres choix sont possibles. Tout dépend de l'information disponible pour un problème donné. Par exemple, il est possible que la pente soit connue aux extrémités ou que la fonction recherchée soit périodique. On peut alors se servir de cette information pour obtenir les deux équations manquantes. □

Remarque 5.13

Pour effectuer une interpolation à l'aide des splines cubiques, il faut en premier lieu résoudre le système 5.31. Par la suite, on doit déterminer l'intervalle dans lequel se situe le point d'interpolation x et calculer le polynôme dans cet intervalle en utilisant la formule 5.30. □

Remarque 5.14

Dans le cas où les abscisses sont équidistantes, c'est-à-dire:

$$h_i = h \ \ \forall i$$

la matrice du système linéaire 5.31 se trouve simplifiée de beaucoup. En effet, on obtient alors une matrice tridiagonale dont la diagonale principale ne contient que des 2, tandis que les deux autres diagonales sont constituées de coefficients valant 1/2. *Cette matrice ne dépend donc pas de la valeur de h, qui n'affecte que le terme de droite.* □

Exemple 5.11

Soit les 4 points suivants: $(1,1)$, $(2,4)$, $(4,9)$, $(5,11)$. On trouve toute l'information nécessaire au calcul de la spline cubique dans la table suivante.

i	x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	h_i
0	1	1			
			3		1
1	2	4		$-1/6$	
			$5/2$		2
2	4	9		$-1/6$	
			2		1
3	5	11			

La première équation ($i = 1$) du système 5.31 devient:

$$(1/3)f_0'' + 2f_1'' + (2/3)f_2'' = 6(-1/6)$$

et la deuxième équation ($i = 2$) s'écrit:

$$(2/3)f_1'' + 2f_2'' + (1/3)f_3'' = 6(-1/6)$$

Pour obtenir la spline naturelle, on pose $f_0'' = f_3'' = 0$, et on obtient le système:

$$\begin{bmatrix} 2 & 2/3 \\ 2/3 & 2 \end{bmatrix} \begin{bmatrix} f_1'' \\ f_2'' \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

dont la solution est $f_1'' = -0{,}375$ et $f_2'' = -0{,}375$. L'équation de la spline dans le premier intervalle est alors:

$$\begin{aligned} p_1(x) \quad = \quad & 0\frac{(x-2)^3}{6} - 0{,}375\frac{(x-1)^3}{6} \\ & - \left(\frac{1}{1} - \frac{0}{6}\right)(x-2) \\ & + \left(\frac{4}{1} - \frac{(1)(-0{,}375)}{6}\right)(x-1) \end{aligned}$$

que l'on peut simplifier en:

$$p_1(x) = -0{,}0625(x-1)^3 - (x-2) + 4{,}0625(x-1)$$

Ce polynôme n'est défini que dans l'intervalle $[1, 2]$. On peut par exemple l'évaluer en $x = 1{,}5$ pour obtenir $2{,}523\,4375$.

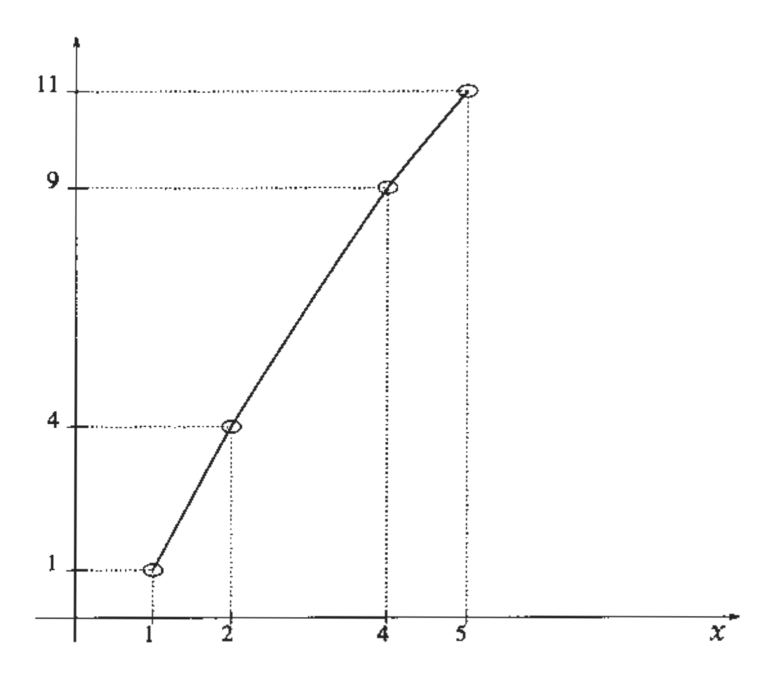


Figure 5.9: Spline passant par 4 points

De même, si on a besoin de la valeur de la spline en $x = 3$, qui est situé dans le deuxième intervalle (soit $[2, 4]$), on peut obtenir l'équation de la spline dans cet intervalle en posant $i = 2$ dans l'équation 5.30. On a alors:

$$\begin{aligned} p_2(x) &= 0{,}375\frac{(x-4)^3}{12} - 0{,}375\frac{(x-2)^3}{12} \\ &\quad - \left(\frac{4}{2} - \frac{(2)(-0{,}375)}{6}\right)(x-4) \\ &\quad + \left(\frac{9}{2} - \frac{(2)(-0{,}375)}{6}\right)(x-2) \end{aligned}$$

c'est-à-dire:

$$p_2(x) = 0{,}031\,25(x-4)^3 - 0{,}031\,25(x-2)^3 - 2{,}125(x-4) + 4{,}625(x-2)$$

La valeur de la spline en $x = 3$ est donc 6,6875. La spline complète est illustrée à la figure 5.9.

● ● ● ● ●

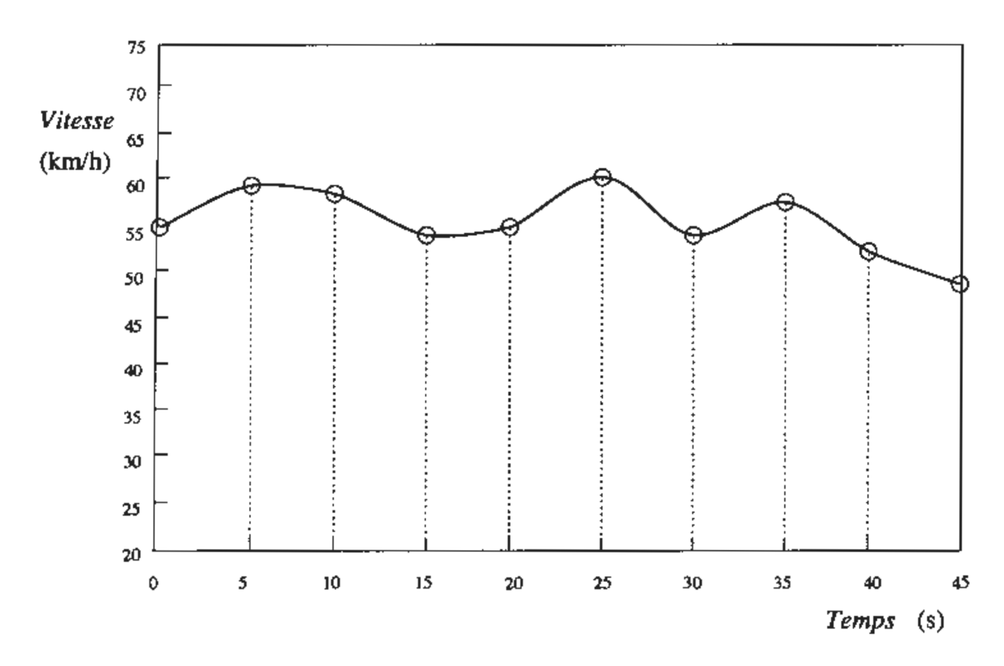


Figure 5.10: Spline cubique associée au problème du véhicule

Exemple 5.12
Si on reprend l'exemple où la vitesse d'un véhicule est mesurée toutes les 5 secondes, on obtient la spline de la figure 5.10. On remarque immédiatement que les fortes oscillations observées à la figure 5.4 avec un polynôme de degré 9 ont maintenant disparu. On peut alors interpoler la valeur de la vitesse du véhicule partout dans l'intervalle $[0, 45]$ sans risque d'obtenir des valeurs aberrantes.

● ● ● ● ●

5.7 Krigeage

Le krigeage, dont l'origine remonte à Krige (réf. [16]) au début des années cinquante, s'est avéré une technique d'interpolation extrêmement puissante à la suite des travaux de Matheron (réf. [19]). Le krigeage généralise un certain nombre de méthodes d'interpolation que nous avons vues, mais son principal avantage est de s'étendre facilement aux cas bidimensionnels et

tridimensionnels. Nous verrons en fait que faire du krigeage en 1, 2, 3 ou plus de 3 dimensions est tout aussi facile.

Il est possible, par une approche simple, de mettre en évidence certaines propriétés du krigeage ou plus précisément du *krigeage dual*. Nous nous inspirerons à cet égard de l'article de Trochu (réf. [23]), qui contient de plus une excellente revue de la littérature ainsi que des applications en ingénierie [1].

Le krigeage a été introduit principalement pour répondre aux besoins de la prospection minière. Il servait notamment à reconstituer la position d'un filon de minerai à partir de concentrations obtenues par forage. Il y a donc un aspect statistique inhérent à ce problème, mais nous ne l'abordons pas ici. Nous nous limitons à présenter le krigeage comme une technique d'interpolation.

Nous abordons en premier lieu l'interpolation d'une courbe plane de forme $y = f(x)$, un problème que nous connaissons bien. Soit les n points $((x_i, f(x_i)$ pour $i = 1, 2, 3, \cdots, n)$. Considérons le système linéaire suivant:

$$\begin{bmatrix} K_{11} & K_{12} & K_{13} & \cdots & K_{1n} & 1 & x_1 \\ K_{21} & K_{22} & K_{23} & \cdots & K_{2n} & 1 & x_2 \\ K_{31} & K_{32} & K_{33} & \cdots & K_{3n} & 1 & x_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ K_{n1} & K_{n2} & K_{n3} & \cdots & K_{nn} & 1 & x_n \\ 1 & 1 & 1 & \cdots & 1 & 0 & 0 \\ x_1 & x_2 & x_3 & \cdots & x_n & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_n \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_n) \\ 0 \\ 0 \end{bmatrix} \tag{5.32}$$

qui peut être représenté sous forme matricielle:

$$\begin{bmatrix} K & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \vec{\alpha} \\ \vec{a} \end{bmatrix} = \begin{bmatrix} \vec{f} \\ \vec{0} \end{bmatrix} \tag{5.33}$$

Les matrices K et A ainsi que les vecteurs $\vec{\alpha}$, $\vec{f}$ et $\vec{a}$ sont définis par la comparaison des relations 5.32 et 5.33. La matrice A dépend des coordonnées des points de collocation, tandis que les éléments de la matrice K sont donnés typiquement par une relation de la forme:

$$K_{ij} = g(|x_i - x_j|) \tag{5.34}$$

[1] M. F. Trochu est professeur au Département de génie mécanique de l'École Polytechnique de Montréal.

On obtient les coefficients K_{ij} à partir d'une fonction g qui varie selon la distance entre les abscisses x_i et x_j. Le choix de la fonction g détermine les propriétés de la courbe de krigeage.

Considérons maintenant la fonction:

$$u(x) = \sum_{j=1}^{n} \alpha_j g(|x - x_j|) + a_1 + a_2 x \tag{5.35}$$

Si a_1, a_2 et les α_j sont solutions du système 5.32, la fonction $u(x)$ passe par les n points d'interpolation donnés. En effet, la i^e équation du système se lit:

$$\sum_{j=1}^{n} \alpha_j K_{ij} + a_1 + a_2 x_i = f(x_i)$$

qui s'écrit également:

$$\sum_{j=1}^{n} \alpha_j g(|x_i - x_j|) + a_1 + a_2 x_i = f(x_i)$$

Cela signifie que:

$$u(x_i) = f(x_i)$$

Les deux dernières équations du système 5.32 sont tout simplement:

$$\sum_{j=1}^{n} \alpha_j = 0 \tag{5.36}$$

et

$$\sum_{j=1}^{n} \alpha_j x_j = 0 \tag{5.37}$$

qui traduisent des conditions de non-biais de la fonction $u(x)$, faisant ici référence à l'aspect statistique du krigeage.

La fonction $u(x)$ se décompose en deux parties. On appelle *dérive* la partie:

$$a_1 + a_2 x \tag{5.38}$$

qui peut également être un polynôme de degré k. Il faut alors modifier légèrement le système 5.32 en ajoutant les lignes et les colonnes contenant les

différentes puissances des points x_i. Le système prend la forme:

$$\begin{bmatrix} K_{11} & \cdots & K_{1n} & 1 & x_1 & (x_1)^2 & \cdots & (x_1)^k \\ K_{21} & \cdots & K_{2n} & 1 & x_2 & (x_2)^2 & \cdots & (x_2)^k \\ K_{31} & \cdots & K_{3n} & 1 & x_3 & (x_3)^2 & \cdots & (x_3)^k \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ K_{n1} & \cdots & K_{nn} & 1 & x_n & (x_n)^2 & \cdots & (x_n)^k \\ 1 & \cdots & 1 & 0 & 0 & 0 & \cdots & 0 \\ x_1 & \cdots & x_n & 0 & 0 & 0 & \cdots & 0 \\ (x_1)^2 & \cdots & (x_n)^2 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ (x_1)^k & \cdots & (x_n)^k & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_n \\ a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_{k+1} \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_n) \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{5.39}$$

La fonction:

$$\sum_{j=1}^{n} \alpha_j g(|x - x_j|) \tag{5.40}$$

est appelée *fluctuation aléatoire*. La dérive peut s'interpréter comme une première approximation du comportement général de la fonction $f(x)$ à interpoler. La fluctuation aléatoire est une correction de la dérive permettant à la courbe de passer par les points d'interpolation. Dans la figure 5.11, les points sont plus ou moins disséminés autour d'une droite. Le choix d'une dérive linéaire paraît alors approprié. Dans d'autres circonstances, une dérive constante ou parabolique pourrait donner de meilleurs résultats.

Les remarques suivantes sur la nature du système 5.32 joueront un rôle important plus loin.

Remarque 5.15
Le choix de la fonction $g(h)$ n'est pas totalement arbitraire; il est régi par des règles que nous ne précisons pas ici. De plus, si on remplace la fonction g par cg, où c est une constante, la matrice K est multipliée par la même constante. Il est facile de voir que la solution du système 5.32 est alors $[\vec{\alpha}/c \;\; \vec{a}]^T$. On en conclut que lorsqu'on multiplie la fonction g par une constante le vecteur solution $\vec{\alpha}$ est divisé par la même constante et le vecteur $\vec{a}$ reste inchangé.

On constate en outre que la fonction $u(x)$ de la formule 5.35 reste inchangée puisque la fonction g est multipliée par c et que les coefficients α_j sont divisés par c. On en conclut que *la courbe de krigeage 5.35 est inchangée lorsque la fonction g est multipliée par une constante.* □

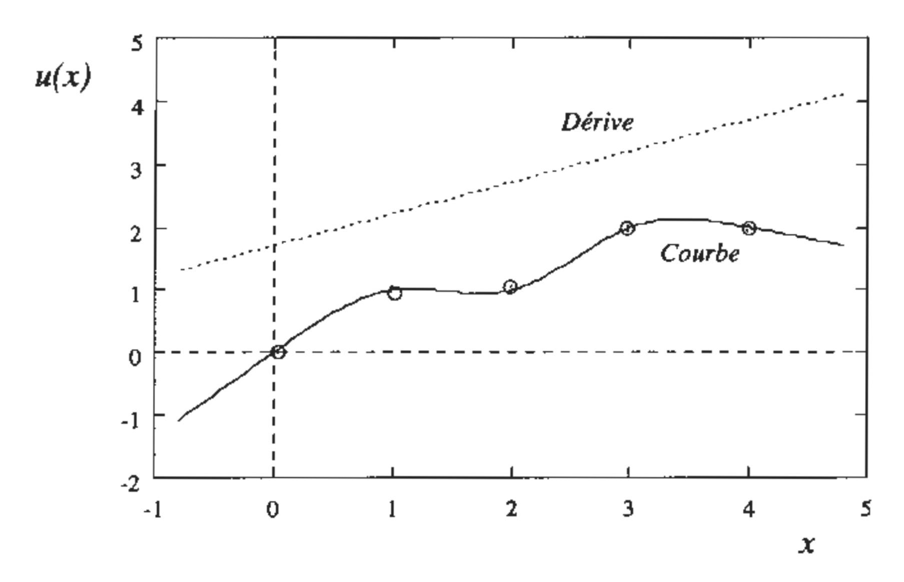


Figure 5.11: Dérive et courbe de krigeage

Remarque 5.16
Si on ajoute une constante c à chaque coefficient de la matrice K (ce qui revient à remplacer g(h) par g(h) + c), la solution du système 5.32 reste inchangée. En effet, la i^{e} équation du système devient:

$$\begin{aligned}
(K_{i1}+c)\alpha_1 + (K_{i2}+c)\alpha_2 + \cdots + (K_{in}+c)\alpha_n + a_1 + a_2 x_i &= f(x_i)\\
\sum_{j=1}^{n} K_{ij}\alpha_j + c\sum_{j=1}^{n}\alpha_j + a_1 + a_2 x_i &= f(x_i)\\
\sum_{j=1}^{n} K_{ij}\alpha_j + a_1 + a_2 x_i &= f(x_i)
\end{aligned}$$

en vertu de l'équation 5.36. □

Le choix de la fonction g n'est pas tout à fait arbitraire et certaines fonctions g sont plus intéressantes que d'autres. Commençons par le choix le plus simple:

$$g(h) = h \tag{5.41}$$

ce qui entraîne que:

$$K_{ij} = g(|x_i - x_j|) = |x_i - x_j|$$

La fonction $u(x)$ s'écrit dans ce cas:

$$u(x) = \sum_{j=1}^{n} \alpha_j |x - x_j| + a_1 + a_2 x$$

La fonction $|x - x_j|$ est linéaire par morceaux et non dérivable au point x_j. La dérive étant également linéaire, on en déduit que $u(x)$ est une interpolation linéaire par morceaux puisque, par construction, $u(x)$ passe par tous les points de collocation. On appelle également cette courbe une *spline linéaire*.

Exemple 5.13
Soit les 5 points suivants: $(0,0)$, $(1,1)$, $(2,1)$, $(3,2)$ et (4,2). Si on utilise la fonction 5.41, le système linéaire 5.32 est de dimension $(5+2)$ et s'écrit:

$$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 1 & 0 \\ 1 & 0 & 1 & 2 & 3 & 1 & 1 \\ 2 & 1 & 0 & 1 & 2 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 & 1 & 3 \\ 4 & 3 & 2 & 1 & 0 & 1 & 4 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 3 & 4 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 2 \\ 0 \\ 0 \end{bmatrix}$$

dont la solution, obtenue par décomposition LU, est le vecteur:

$$[\frac{1}{4} \; -\frac{1}{2} \; \frac{1}{2} \; -\frac{1}{2} \; \frac{1}{4} \; 0 \; \frac{1}{2}]^T$$

La fonction $u(x)$ s'écrit alors:

$$u(x) = \frac{1}{4}|x-0| - \frac{1}{2}|x-1| + \frac{1}{2}|x-2| - \frac{1}{2}|x-3| + \frac{1}{4}|x-4| + 0 + \frac{1}{2}x$$

et est illustrée à la figure 5.12. On remarque que cette fonction est bien linéaire par morceaux et passe par tous les points de collocation, d'où cet aspect en dents de scie. De plus, pour obtenir l'équation de cette spline

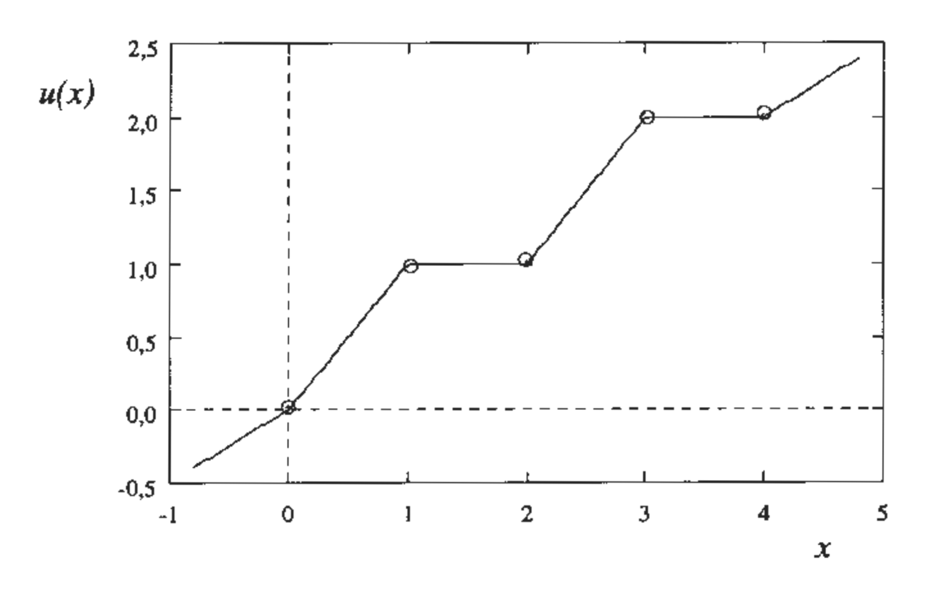


Figure 5.12: Krigeage linéaire

linéaire à l'aide des techniques classiques, il faudrait définir la courbe pour chaque sous-intervalle. Ici, on profite d'une seule expression valide dans tout le domaine. En outre, l'extrapolation est permise.

● ● ● ● ●

Le deuxième cas intéressant correspond au choix:

$$g(h) = h^3 \tag{5.42}$$

Les coefficients de la matrice K sont alors des fonctions du cube de la distance entre les abscisses d'interpolation, c'est-à-dire:

$$K_{ij} = g(|x_i - x_j|) = |x_i - x_j|^3$$

et la fonction $u(x)$ correspondante fait intervenir les fonctions:

$$g_j(x) = g(|x - x_j|) = |x - x_j|^3$$

Chaque fonction $g_j(x)$ est deux fois différentiable au point x_j. En effet:

$$g_j(x) = \begin{cases} -(x - x_j)^3 & \text{si} \quad x < x_j \\ (x - x_j)^3 & \text{si} \quad x > x_j \end{cases}$$

Les limites à gauche et à droite en x_j tendent vers 0, ce qui signifie que cette fonction est continue en x_j. De même:

$$g_j'(x) = \begin{cases} -3(x - x_j)^2 & \text{si} \quad x < x_j \\ 3(x - x_j)^2 & \text{si} \quad x > x_j \end{cases}$$

et les limites à gauche et à droite tendent encore vers 0. La fonction $g_j(x)$ est donc dérivable une première fois. De plus:

$$g_j''(x) = \begin{cases} -6(x - x_j) & \text{si} \quad x < x_j \\ 6(x - x_j) & \text{si} \quad x > x_j \end{cases}$$

et

$$g_j'''(x) = \begin{cases} -6 & \text{si} \quad x < x_j \\ 6 & \text{si} \quad x > x_j \end{cases}$$

On constate donc que la dérivée seconde est également continue, mais pas la dérivée troisième puisque les limites à gauche et à droite valent respectivement -6 et 6. La fonction $u(x)$ est donc deux fois différentiable et est de degré 3 partout, ce qui démontre qu'il s'agit d'une spline cubique. Cette spline est implicitement naturelle puisque:

$$\begin{aligned} u''(x_1) &= \sum_{j=1}^{n} \alpha_j g_j''(x_1) \\ &= -\sum_{j=1}^{n} \alpha_j 6(x_1 - x_j) \\ &= -6x_1 \sum_{j=1}^{n} \alpha_j + 6 \sum_{j=1}^{n} \alpha_j x_j \\ &= 0 \end{aligned}$$

en vertu des relations 5.36 et 5.37. On obtiendrait un résultat similaire avec $u''(x_n)$.

Exemple 5.14
Si on considère les mêmes points que dans l'exemple précédent mais avec $g(h) = h^3$, on obtient le système:

$$\begin{bmatrix} 0 & 1 & 8 & 27 & 64 & 1 & 0 \\ 1 & 0 & 1 & 8 & 27 & 1 & 1 \\ 8 & 1 & 0 & 1 & 8 & 1 & 2 \\ 27 & 8 & 1 & 0 & 1 & 1 & 3 \\ 64 & 27 & 8 & 1 & 0 & 1 & 4 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 3 & 4 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 2 \\ 0 \\ 0 \end{bmatrix}$$

La fonction $u(x)$ correspondante est illustrée à la figure 5.13 et correspond bien à une spline cubique. L'équation de cette spline est:

$$\begin{aligned} u(x) &= -0{,}178\,75|x-0|^3 + 0{,}571\,43|x-1|^3 - 0{,}785\,71|x-2|^3 \\ &\quad + 0{,}571\,43|x-3|^3 - 0{,}178\,57|x-4|^3 + 1{,}7143 + 0{,}5x \end{aligned}$$

On remarque que l'expression de la dérive n'est pas la même que dans le cas de $g(h) = h$. De plus, l'extrapolation de part et d'autre de l'intervalle $[0, 4]$ se fait de manière linéaire, ce qui correspond bien à une spline naturelle.
Il est également intéressant de souligner qu'une seule équation exprime la courbe complète, contrairement au système d'équations par sous-intervalles décrit à la section 5.6 (voir la formule 5.30).

● ● ● ● ●

Remarque 5.17
Une autre différence importante marque cette nouvelle façon d'obtenir l'équation de la spline. À la section 5.6, un système linéaire tridiagonal était nécessaire au calcul de la spline. Si on recourt au krigeage, le système linéaire correspondant n'est plus tridiagonal. □

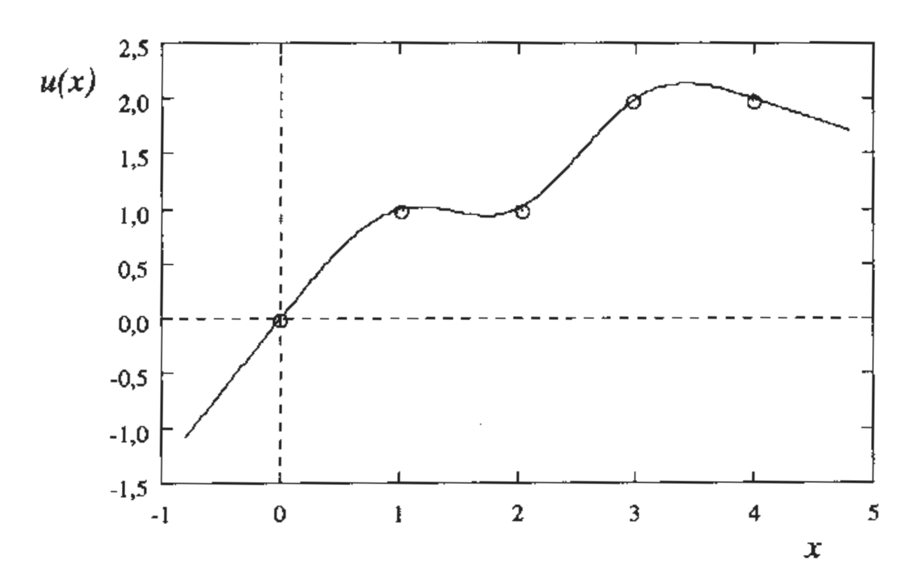


Figure 5.13: Krigeage cubique

Exemple 5.15
On a vu que la matrice K est pleine, c'est-à-dire que la majeure partie des coefficients de cette matrice sont non nuls. Cela tient au fait que ces coefficients sont définis par:

$$K_{ij} = g(|x_i - x_j|)$$

et que les différents choix de fonction $g(h)$ faits jusqu'à maintenant ne permettent d'annuler les éléments K_{ij} que sur la diagonale.

Il est intuitivement clair que l'interpolation d'une fonction en un point x dépend plus fortement des points de collocation situés dans le voisinage immédiat de x que des points plus éloignés. Pour tenir compte de cette observation, il est possible d'introduire une distance d'influence d au-delà de laquelle la fonction $g(h)$ s'annule, ce qui permet de réduire l'influence des points de collocation situés à une distance supérieure à d. Par exemple (voir Trochu, réf. [23]), on peut obtenir une courbe similaire aux splines cubiques en définissant:

$$g(h) = \begin{cases} 1 - (\frac{h}{d})^3 & \text{si } 0 \leq h \leq d \\ 0 & \text{si } h > d \end{cases}$$

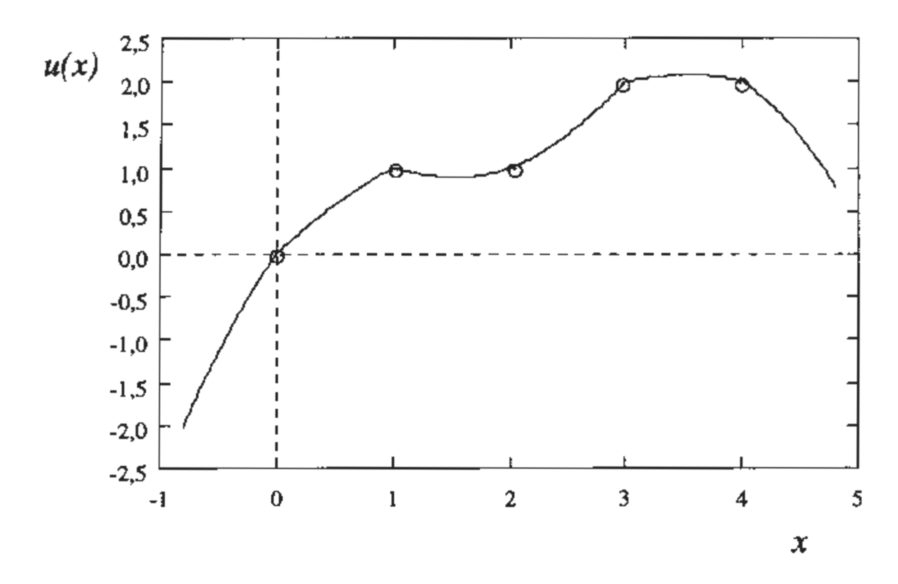


Figure 5.14: Krigeage cubique: distance d'influence $d = 3$

Cette définition assure la continuité de la fonction g. On pourrait également modifier légèrement cette fonction pour la rendre différentiable, sans changer les propriétés de la courbe. Si la distance d est très grande, on retrouvera une spline cubique comme précédemment. Par contre, si on réduit la valeur de d, la courbe aura tendance à osciller plus fortement. Cela se voit sur la figure 5.14, où on a utilisé les mêmes points de collocation que dans les exemples précédents mais en introduisant une distance d'influence $d = 3$. Le résultat est encore plus probant avec $d = 0{,}5$, comme en témoigne la figure 5.15 où la distance d'influence est nettement perceptible.

• • • • •

Remarque 5.18
Les systèmes linéaires de krigeage des exemples précédents sont quelque peu particuliers. En effet, la diagonale principale des matrices de krigeage est *a priori* nulle. Il faut donc se montrer prudent au moment de la résolution par décomposition LU. La recherche d'un pivot non nul peut s'avérer nécessaire. On peut également faciliter le traitement numérique du système en modifiant la fonction $g(h)$, par exemple en lui ajoutant une constante, sans modifier la courbe de krigeage. □

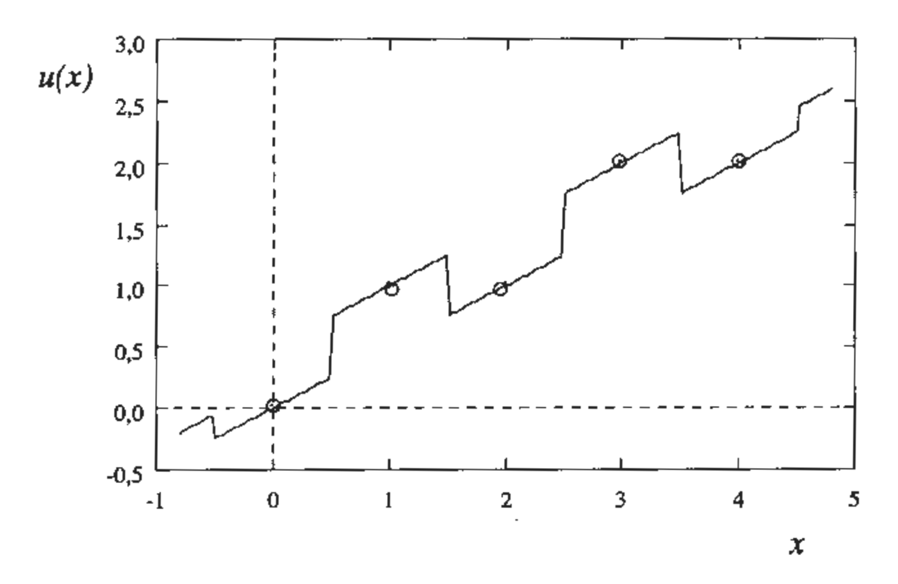


Figure 5.15: Krigeage cubique: distance d'influence $d = 0{,}5$

5.7.1 Effet pépite

Lorsqu'on fait de l'interpolation à partir de mesures expérimentales, il est quelquefois utile d'avoir la possibilité d'éliminer une donnée qui paraît aberrante. Cette donnée peut provenir d'une erreur de mesure par exemple. Initialement, dans la prospection minière, on a qualifié la présence de ces données marginales d'*effet pépite*. Cette expression traduisait l'empressement de certains analystes miniers à conclure trop vite à une forte concentration d'or dans le voisinage immédiat d'une pépite d'or isolée.

Une façon de contourner cette difficulté est de pondérer les mesures expérimentales en y attachant un poids variable suivant la fiabilité de la mesure. Plus précisément, on attache un poids d'autant plus grand que la variance statistique de l'erreur liée à cette mesnre est petite. On voit encore ici poindre l'aspect statistique du krigeage.

Sur le plan pratique, cela se fait en modifiant très légèrement la matrice K du système de krigeage 5.32. Il suffit en fait de modifier la diagonale de

la matrice K et de considérer le système linéaire:

$$\begin{bmatrix} K_{11}+w_1 & K_{12} & \cdots & K_{1n} & 1 & x_1 \\ K_{21} & K_{22}+w_2 & \cdots & K_{2n} & 1 & x_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ K_{n1} & K_{n2} & \cdots & K_{nn}+w_n & 1 & x_n \\ 1 & 1 & \cdots & 1 & 0 & 0 \\ x_1 & x_2 & \cdots & x_n & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \\ 0 \\ 0 \end{bmatrix} \tag{5.43}$$

que l'on peut représenter sous forme matricielle:

$$\begin{bmatrix} K+D & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \vec{\alpha} \\ \vec{a} \end{bmatrix} = \begin{bmatrix} \vec{f} \\ \vec{0} \end{bmatrix} \tag{5.44}$$

La matrice diagonale:

$$D = \begin{bmatrix} w_1 & 0 & 0 & \cdots & 0 \\ 0 & w_2 & 0 & \cdots & \vdots \\ 0 & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & 0 & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & w_n \end{bmatrix}$$

exprime en quelque sorte le degré de fiabilité attaché à chaque mesure. *Plus w_i est grand, moins on tient compte de cette mesure dans le calcul de la fonction de krigeage.* Sur le plan statistique, w_i est proportionnel à la variance de l'erreur sur la i^e mesure.

Remarque 5.19
Pour mettre en évidence un effet pépite, il n'est nécessaire de modifier que la diagonale de la matrice K du système de krigeage. Le choix de $g(h)$ et l'expression de $u(x)$ restent inchangés. □

On remarque immédiatement que si $w_i \neq 0$:

$$u(x_i) \neq f(x_i)$$

ce qui entraîne que la fonction de krigeage ne passe pas par le point $(x_i, f(x_i))$. En effet, l'équation correspondante du système 5.43 devient:

$$K_{i1}\alpha_1 + \cdots + K_{i\,i-1}\alpha_{i-1} + (K_{ii} + w_i)\alpha_i + K_{i\,i+1}\alpha_{i+1} + \cdots + K_{in}\alpha_n = f(x_i)$$

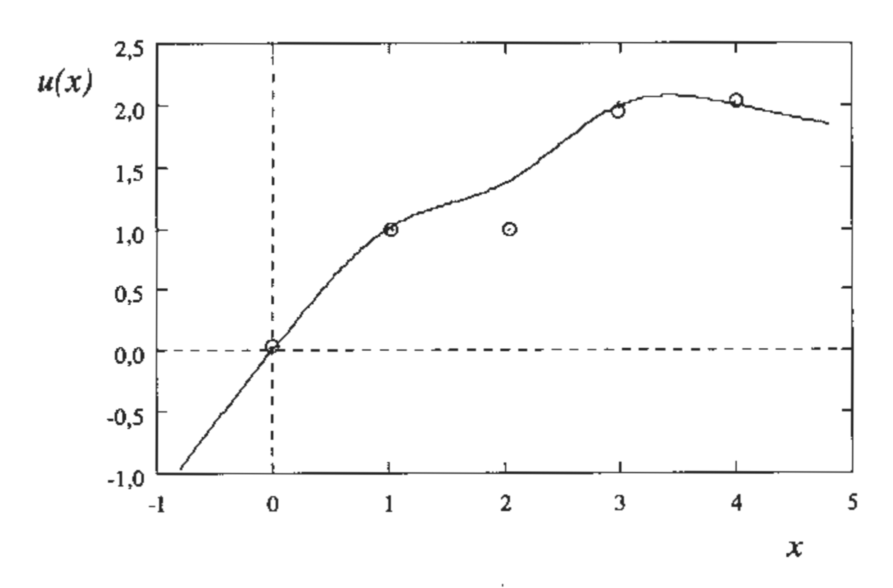


Figure 5.16: Krigeage cubique: effet pépite

qui devient:

$$(K_{i1}\alpha_1 + \cdots + K_{i\,i-1}\alpha_{i-1} + K_{ii}\alpha_i + K_{i\,i+1}\alpha_{i+1} + \cdots + K_{in}\alpha_n) + w_i\alpha_i = f(x_i)$$

ou encore

$$u(x_i) + w_i\alpha_i = f(x_i)$$

Plus w_i est grand, plus l'écart entre $u(x_i)$ et $f(x_i)$ risque d'être grand.

Exemple 5.16
Reprenons les 5 points de l'exemple précédent avec $g(h) = h^3$, mais en introduisant cette fois un effet pépite au troisième point ($w_3 = 1$, $w_i = 0$ $\forall i \neq 3$). On obtient la spline cubique de la figure 5.16. On remarque que la spline ne passe plus par le troisième point.

● ● ● ● ●

On peut se demander ce qui arrive lorsque l'effet pépite devient très important. Pour répondre à cette question, il est nécessaire de remplacer la matrice D par βD et de faire tendre β vers l'infini. De cette manière, l'importance relative des w_i est la même quelle que soit la valeur de β.

Le système 5.44 s'écrit:

$$\begin{aligned}(K+\beta D)\vec{\alpha}+A\vec{a} &= \vec{f}\\ A^T\vec{\alpha} &= \vec{0}\end{aligned}$$

qui, lorsqu'on isole $\vec{\alpha}$ dans la première équation et qu'on remplace cette variable dans la deuxième équation, devient:

$$A^T(K+\beta D)^{-1}(\vec{f}-A\vec{a})=0$$

ou encore

$$A^T(K+\beta D)^{-1}A\vec{a}=A^T(K+\beta D)^{-1}\vec{f} \tag{5.45}$$

Théorème 5.6

Lorsque $\beta \to \infty$, le système 5.45 tend vers le système:

$$A^TD^{-1}A\vec{a}=A^TD^{-1}\vec{f} \tag{5.46}$$

qui correspond à un problème de moindres carrés pondérés. En particulier, si la matrice D est la matrice identité (ce qui correspond à une pondération égale pour tous les points), on trouve:

$$A^TA\vec{a}=A^T\vec{f} \tag{5.47}$$

de telle sorte que la dérive a_1+a_2x n'est rien d'autre que la droite de moindres carrés (voir Burden et Faires, réf. [2]).

Démonstration (facultative):

On considère en premier lieu la matrice $(K+\beta D)$, qui devient:

$$(K+\beta D)=\beta D(I+\frac{1}{\beta}D^{-1}K)$$

Si β est suffisamment grand:

$$||\frac{1}{\beta}D^{-1}K||_\infty<1$$

ce qui entraîne que le rayon spectral de cette matrice est inférieur à 1 et que la matrice est convergente. On a alors:

$$(I+\frac{1}{\beta}D^{-1}K)^{-1}=I-(\frac{1}{\beta}D^{-1}K)+(\frac{1}{\beta}D^{-1}K)^2-(\frac{1}{\beta}D^{-1}K)^3+\cdots$$

ou plus précisément:

$$\begin{aligned}(K+\beta D)^{-1} &= (I+\tfrac{1}{\beta}D^{-1}K)^{-1}\tfrac{1}{\beta}D^{-1} \\ &= \tfrac{1}{\beta}(I+\tfrac{1}{\beta}D^{-1}K)^{-1}D^{-1} \\ &= \tfrac{1}{\beta}\left[I-\tfrac{1}{\beta}D^{-1}K+(\tfrac{1}{\beta}D^{-1}K)^2-(\tfrac{1}{\beta}D^{-1}K)^3+\cdots\right]D^{-1} \\ &= \tfrac{1}{\beta}\left[D^{-1}-\tfrac{1}{\beta}D^{-1}KD^{-1}+(\tfrac{1}{\beta}D^{-1}K)^2D^{-1}+\cdots\right]\end{aligned}$$

Le système de krigeage 5.45 devient alors:

$$\begin{aligned}&\tfrac{1}{\beta}A^T\left[D^{-1}-\tfrac{1}{\beta}D^{-1}KD^{-1}+(\tfrac{1}{\beta}D^{-1}K)^2D^{-1}+\cdots\right]A\vec{a} \\ =\ &\tfrac{1}{\beta}A^T\left[D^{-1}-\tfrac{1}{\beta}D^{-1}KD^{-1}+(\tfrac{1}{\beta}D^{-1}K)^2D^{-1}+\cdots\right]\vec{f}\end{aligned}$$

où on peut simplifier un coefficient $1/\beta$ de chaque côté. En faisant tendre β vers l'infini, on trouve immédiatement:

$$A^TD^{-1}A\vec{a} = A^TD^{-1}\vec{f}$$

qui est le résultat recherché. □

En pratique, il n'est pas nécessaire de faire tendre β vers l'infini. Il suffit en effet de prendre des poids w_i très grands.

Exemple 5.17
En reprenant les points de l'exemple précédent mais avec $w_i = 10^5\ \ \forall i$, on trouve la droite de la figure 5.17. Il s'agit bien de la droite de moindres carrés.

● ● ● ● ●

5.7.2 Courbes paramétrées

Il est parfois utile de construire des courbes paramétrées. Cela permet, par exemple, de construire des courbes fermées (comme un cercle ou une ellipse)

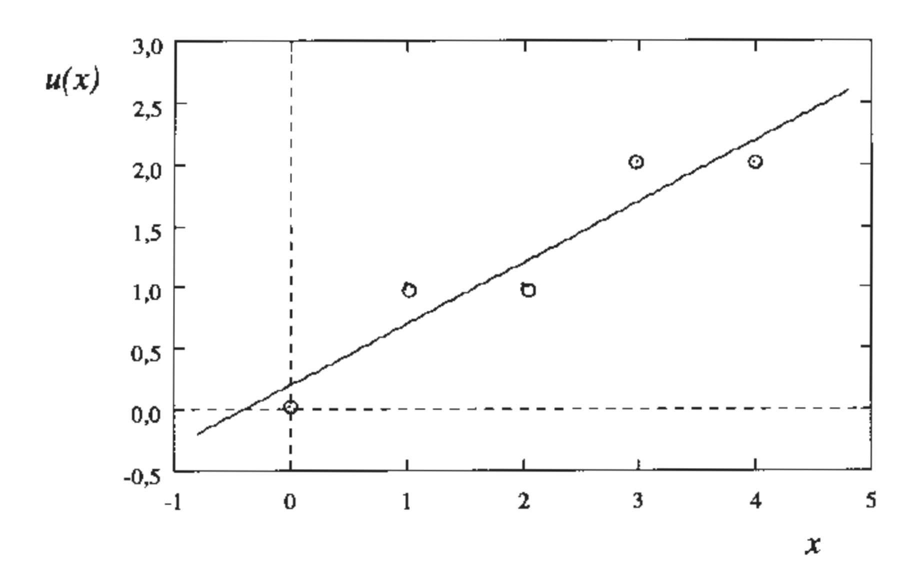


Figure 5.17: Krigeage cubique: effet pépite avec $w_i = 10^5$

ou plus généralement des courbes qui sont en intersection avec elles-mêmes (telle une boucle). On peut ainsi, à l'aide de quelques points donnés dans l'espace, construire des trajectoires complexes. La paramétrisation d'une courbe de l'espace s'obtient par une équation de la forme:

$$\vec{\gamma}(t) = (\gamma_1(t), \gamma_2(t), \gamma_3(t)), \quad t \in [a, b]$$

On décrit la courbe en faisant varier le paramètre t entre a et b. La stratégie de krigeage reste sensiblement la même dans ce cas. En effet, pour construire une courbe paramétrée passant par les points $((x_1^i, x_2^i, x_3^i)$ pour $i = 1, 2, \cdots, n)$, il faut d'abord construire la suite t_i des valeurs du paramètre t de telle sorte que:

$$\vec{\gamma}(t_i) = (x_1^i, x_2^i, x_3^i)$$

Le choix le plus simple est bien sûr:

$$t_i = i$$

mais il ne tient aucun compte des distances respectives entre les points (x_1^i, x_2^i, x_3^i). Selon Trochu (réf. [23]), un choix plus judicieux consiste à prendre $t_1 = 0$ et:

$$t_{i+1} = t_i + ||\vec{x}^{i+1} - \vec{x}^i||_e \text{ pour } i \geq 1 \tag{5.48}$$

La distance entre les valeurs du paramètre t est ainsi variable et dépend directement de la distance entre les points d'interpolation. Il suffit ensuite de résoudre les systèmes linéaires suivants:

$$\begin{bmatrix} K_{11} & K_{12} & K_{13} & \cdots & K_{1n} & 1 & t_1 \\ K_{21} & K_{22} & K_{23} & \cdots & K_{2n} & 1 & t_2 \\ K_{31} & K_{32} & K_{33} & \cdots & K_{3n} & 1 & t_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ K_{n1} & K_{n2} & K_{n3} & \cdots & K_{nn} & 1 & t_n \\ 1 & 1 & 1 & \cdots & 1 & 0 & 0 \\ t_1 & t_2 & t_3 & \cdots & t_n & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1^1 \\ \alpha_2^1 \\ \alpha_3^1 \\ \vdots \\ \alpha_n^1 \\ a_1^1 \\ a_2^1 \end{bmatrix} = \begin{bmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \\ \vdots \\ x_1^n \\ 0 \\ 0 \end{bmatrix} \tag{5.49}$$

$$\begin{bmatrix} K_{11} & K_{12} & K_{13} & \cdots & K_{1n} & 1 & t_1 \\ K_{21} & K_{22} & K_{23} & \cdots & K_{2n} & 1 & t_2 \\ K_{31} & K_{32} & K_{33} & \cdots & K_{3n} & 1 & t_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ K_{n1} & K_{n2} & K_{n3} & \cdots & K_{nn} & 1 & t_n \\ 1 & 1 & 1 & \cdots & 1 & 0 & 0 \\ t_1 & t_2 & t_3 & \cdots & t_n & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1^2 \\ \alpha_2^2 \\ \alpha_3^2 \\ \vdots \\ \alpha_n^2 \\ a_1^2 \\ a_2^2 \end{bmatrix} = \begin{bmatrix} x_2^1 \\ x_2^2 \\ x_2^3 \\ \vdots \\ x_2^n \\ 0 \\ 0 \end{bmatrix} \tag{5.50}$$

$$\begin{bmatrix} K_{11} & K_{12} & K_{13} & \cdots & K_{1n} & 1 & t_1 \\ K_{21} & K_{22} & K_{23} & \cdots & K_{2n} & 1 & t_2 \\ K_{31} & K_{32} & K_{33} & \cdots & K_{3n} & 1 & t_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ K_{n1} & K_{n2} & K_{n3} & \cdots & K_{nn} & 1 & t_n \\ 1 & 1 & 1 & \cdots & 1 & 0 & 0 \\ t_1 & t_2 & t_3 & \cdots & t_n & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1^3 \\ \alpha_2^3 \\ \alpha_3^3 \\ \vdots \\ \alpha_n^3 \\ a_1^3 \\ a_2^3 \end{bmatrix} = \begin{bmatrix} x_3^1 \\ x_3^2 \\ x_3^3 \\ \vdots \\ x_3^n \\ 0 \\ 0 \end{bmatrix} \tag{5.51}$$

La courbe paramétrée est alors donnée par:

$$\gamma_1(t) = \sum_{j=1}^{n} \alpha_j^1 g(|t - t_j|) + a_1^1 + a_2^1 t \tag{5.52}$$

$$\gamma_2(t) = \sum_{j=1}^{n} \alpha_j^2 g(|t - t_j|) + a_1^2 + a_2^2 t \tag{5.53}$$

$$\gamma_3(t) = \sum_{j=1}^{n} \alpha_j^3 g(|t - t_j|) + a_1^3 + a_2^3 t \tag{5.54}$$

Ici encore, on prouve facilement que:

$$\gamma_1(t_i) = x_1^i, \quad \gamma_2(t_i) = x_2^i \text{ et } \gamma_3(t_i) = x_3^i \text{ pour } 1 \leq i \leq n$$

Remarque 5.20
Les 3 systèmes linéaires requis pour le krigeage paramétré ont tous la même matrice. Seul le membre de droite change. Il est alors important de n'effectuer qu'une seule décomposition LU, suivie de 3 remontées et descentes triangulaires. Cela fait considérablement diminuer le temps de calcul. □

Exemple 5.18
On donne les 12 points suivants de l'espace à 3 dimensions.

x_1^i	x_2^i	x_3^i
0,0	0,0	0,0
1,0	0,0	1,0
1,0	1,0	2,0
0,0	1,0	3,0
0,0	0,0	4,0
1,0	0,0	5,0
1,0	1,0	6,0
0,0	1,0	7,0
0,0	0,0	8,0
1,0	0,0	9,0
1,0	1,0	10,0
0,0	1,0	11,0

On veut construire la courbe paramétrée passant par ces 12 points. Le choix de la fonction $g(h)$ est, comme toujours, primordial. En choisissant d'abord une interpolation linéaire ($g(h) = h$), on obtient la courbe paramétrée de la figure 5.18, qui est peu satisfaisante sur le plan esthétique. En revanche, avec $g(h) = h^3$, on obtient la spirale de la figure 5.19, qui est une spline paramétrée.

● ● ● ● ●

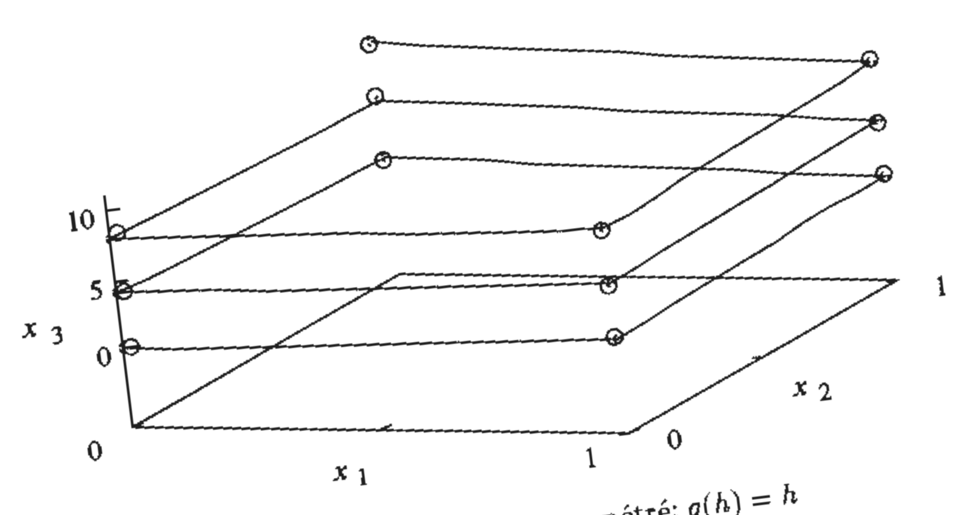


Figure 5.18: Krigeage paramétré: $g(h) = h$

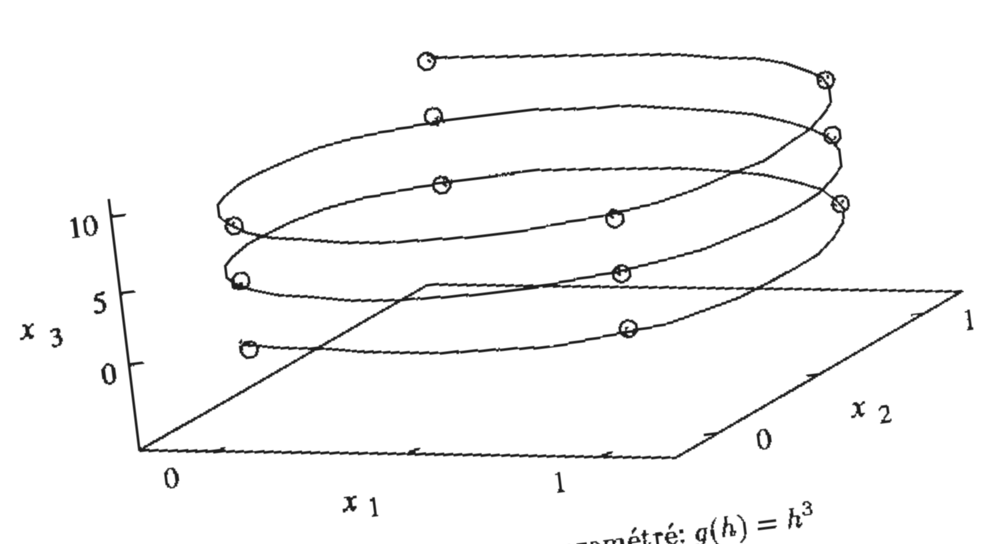


Figure 5.19: Krigeage paramétré: $g(h) = h^3$

5.7.3 Cas multidimensionnel

Le krigeage dual s'étend facilement en plusieurs dimensions. Soit les n points d'interpolation:

$$(x_1^i, x_2^i, x_3^i, f(x_1^i, x_2^i, x_3^i)) \quad \text{pour} \quad i = 1, 2, 3, \cdots, n$$

Le système de krigeage dual devient dans ce cas:

$$\begin{bmatrix} K_{11} & K_{12} & \cdots & K_{1n} & 1 & x_1^1 & x_2^1 & x_3^1 \\ K_{21} & K_{22} & \cdots & K_{2n} & 1 & x_1^2 & x_2^2 & x_3^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ K_{n1} & K_{n2} & \cdots & K_{nn} & 1 & x_1^n & x_2^n & x_3^n \\ 1 & 1 & \cdots & 1 & 0 & 0 & 0 & 0 \\ x_1^1 & x_1^2 & \cdots & x_1^n & 0 & 0 & 0 & 0 \\ x_2^1 & x_2^2 & \cdots & x_2^n & 0 & 0 & 0 & 0 \\ x_3^1 & x_3^2 & \cdots & x_3^n & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} f(x_1^1, x_2^1, x_3^1) \\ f(x_1^2, x_2^2, x_3^2) \\ \vdots \\ f(x_1^n, x_2^n, x_3^n) \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{5.55}$$

et la fonction de krigeage devient:

$$u(\vec{x}) = \sum_{j=1}^{n} \alpha_j g(||\vec{x} - \vec{x}^j||_e) + a_1 + a_2 x_1 + a_3 x_2 + a_4 x_3 \tag{5.56}$$

valable en tout point $\vec{x} = (x_1, x_2, x_3)$. Dans l'expression 5.56, on a remplacé la valeur absolue par la norme euclidienne. En fait, on peut utiliser toute autre norme vectorielle. Il est encore une fois facile de démontrer que:

$$u(x_1^i, x_2^i, x_3^i) = f(x_1^i, x_2^i, x_3^i) \quad \text{pour} \quad i = 1, 2, \cdots, n$$

Remarque 5.21
Le système 5.55 permet le calcul d'une fonction $u(x_1, x_2, x_3)$ de R^3 dans R. On peut également construire des fonctions de R^2 dans R en retirant une dimension du système et en résolvant:

$$\begin{bmatrix} K_{11} & K_{12} & \cdots & K_{1n} & 1 & x_1^1 & x_2^1 \\ K_{21} & K_{22} & \cdots & K_{2n} & 1 & x_1^2 & x_2^2 \\ K_{31} & K_{32} & \cdots & K_{3n} & 1 & x_1^3 & x_2^3 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ K_{n1} & K_{n2} & \cdots & K_{nn} & 1 & x_1^n & x_2^n \\ 1 & 1 & \cdots & 1 & 0 & 0 & 0 \\ x_1^1 & x_1^2 & \cdots & x_1^n & 0 & 0 & 0 \\ x_2^1 & x_2^2 & \cdots & x_2^n & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_n \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} f(x_1^1, x_2^1) \\ f(x_1^2, x_2^2) \\ f(x_1^3, x_2^3) \\ \vdots \\ f(x_1^n, x_2^n) \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{5.57}$$

La fonction $u(x_1, x_2)$ est alors définie par:

$$u(\vec{x}) = \sum_{j=1}^{n} \alpha_j g(||\vec{x} - \vec{x}^j||_e) + a_1 + a_2 x_1 + a_3 x_2 \tag{5.58}$$

Cela montre bien que le krigeage demeure aussi simple en 1 dimension qu'en 2 ou 3 dimensions. □

Remarque 5.22
On peut utiliser l'équation 5.58 pour obtenir des lignes de contour d'une fonction $f(x_1, x_2)$, notamment en topographie. On peut consulter à ce sujet le texte de Trochu (réf. [23]). □

Il reste le choix de la fonction $g(h) = g(||\vec{x} - \vec{x}^j||_e)$. En dimension 1, on a établi que le choix $g(h) = h^3$ conduisait aux splines cubiques (pour une dérive linéaire). On pourrait également montrer que l'équivalent des splines cubiques est obtenu en posant:

$$g(h) = h$$

en dimension 3 et:

$$g(h) = h^2 \ln h$$

en dimension 2.

Exemple 5.19
Soit les 9 points:

x_1^i	x_2^i	$f(x_1^i, x_2^i))$
0,0	0,0	0,0
0,0	0,5	1,0
1,0	0,0	1,0
0,0	1,0	1,5
1,0	0,5	2,0
2,0	0,0	0,25
1,0	1,0	1,0
2,0	0,5	1,0
2,0	1,0	0,0

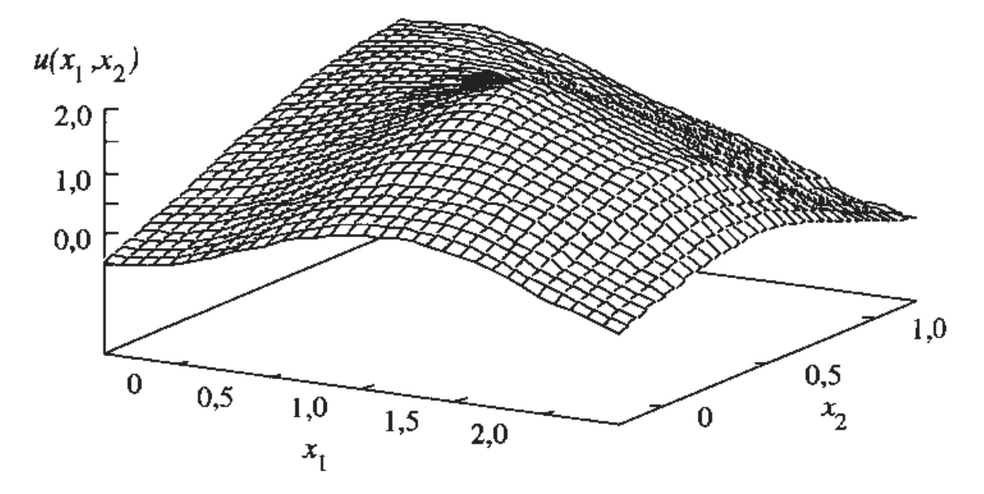


Figure 5.20: Krigeage bidimensionnel: $g(h) = h^2 \ln h$

qui définissent une surface et qui nécessitent un krigeage bidimensionnel. Le système linéaire 5.57 est, dans ce cas, de dimension 12 et les coefficients de la matrice K sont donnés par:

$$K_{ij} = g(||\vec{x}^i - \vec{x}^j||_e) = ||\vec{x}^i - \vec{x}^j||_e^2 \ln(||\vec{x}^i - \vec{x}^j||_e)$$

La fonction de krigeage $u(x_1, x_2)$ prend la forme de la surface de la figure 5.20. Cette surface correspond à l'approximation dite de *coque mince* et est l'équivalent bidimensionnel d'une spline cubique.

●●●●●

Cette section n'est qu'une courte introduction au krigeage. La théorie sous-jacente est vaste et beaucoup de travaux d'applications du krigeage sont actuellement en cours. L'article de Duchon (réf. [9]) traite plus en profondeur de la relation entre le krigeage et les splines cubiques. En ce qui concerne l'aspect statistique, l'article de Matheron (réf. [19]) sur les fonctions aléatoires intrinsèques est certes l'un des plus importants sur le sujet.

5.8 Application: courbe des puissances classées

La courbe des puissances classées d'un service d'électricité (par exemple Hydro-Québec) représente la proportion de l'année où la demande d'électricité atteint ou dépasse une puissance donnée (en gigawatts ou GW). Plus la puissance est grande, plus petite est la proportion de l'année où la demande

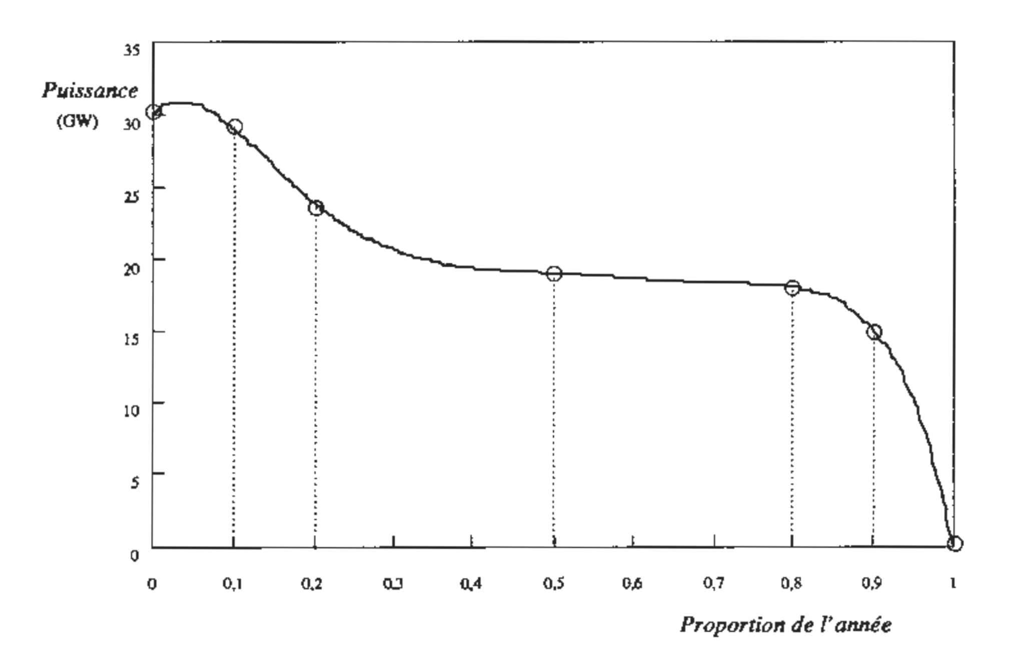


Figure 5.21: Courbe des puissances classées: polynôme de degré 6

dépasse cette valeur. Ainsi, la puissance maximale n'est atteinte que pendant une infime portion de l'année, généralement durant les grands froids de l'hiver. *Cette courbe est par définition décroissante.*

Pour une certaine année de référence, on dispose des données (fictives) suivantes.

Proportion de l'année	Puissance (GW)
0,0	30
0,1	29
0,2	24
0,5	19
0,8	18
0,9	15
1,0	0

La figure 5.21 représente la courbe obtenue par interpolation à l'aide d'un polynôme de degré 6, tandis que la figure 5.22 représente la spline obtenue à partir des mêmes données. Qualitativement, il est évident que la

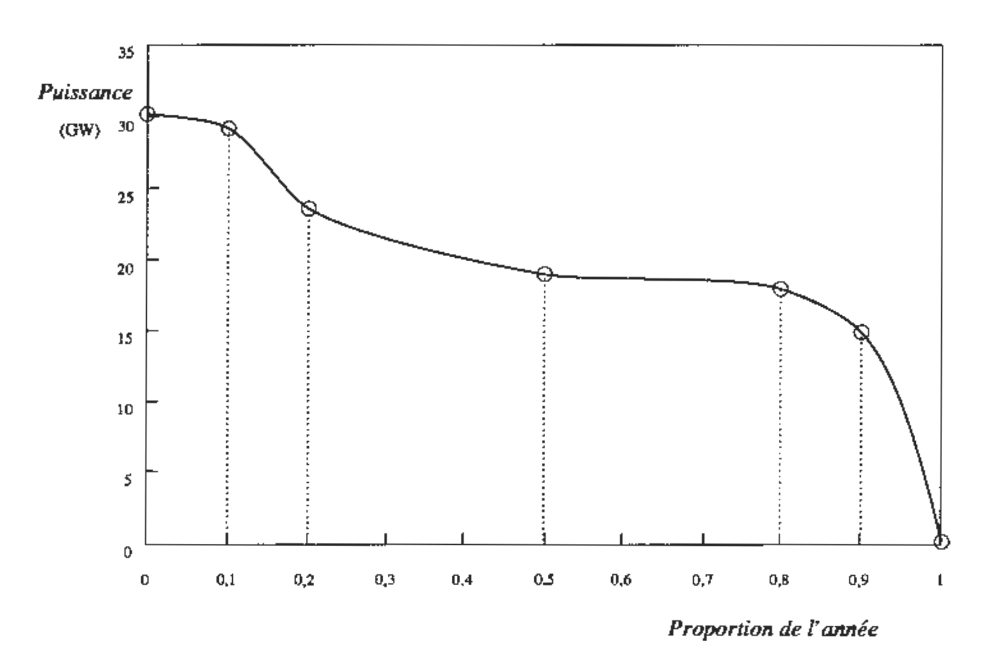


Figure 5.22: Courbe des puissances classées: spline

spline donne de meilleurs résultats puisque la courbe est bien décroissante. Ce n'est pas le cas du polynôme de degré 6, qui est légèrement croissant au début de l'intervalle. Si on interpole ces deux courbes en $x = 0{,}3$, on obtient 20,65 GW à l'aide de l'interpolation de Newton et 20,537 GW au moyen de la spline. On en conclut que pendant 30 % de l'année la puissance demandée dépasse 20,5 GW.

5.9 Exercices

1. Il n'existe pas de polynôme de degré n dont la courbe passe par $(n+2)$ points donnés. Commenter.

2. Obtenir le polynôme de degré 2 passant par les points suivants.

x	$f(x)$
1,0	2,0
2,0	6,0
3,0	12,0

Utiliser la matrice de Vandermonde.

3. Soit les points suivants.

x	$f(x)$
0,0	0,0
1,0	2,0
2,0	36,0
3,0	252,0
4,0	1040,0

a) Obtenir le polynôme de Lagrange passant par les 3 premiers points.

b) Obtenir le polynôme de Lagrange passant par les 4 premiers points. Est-ce possible d'utiliser les calculs faits en a)?

c) Donner l'expression analytique de l'erreur pour les polynômes obtenus en a) et en b).

d) Obtenir des approximations de $f(1{,}5)$ à l'aide des 2 polynômes obtenus en a) et en b).

4. Répondre aux mêmes questions qu'à l'exercice précédent, mais en utilisant la méthode de Newton. Donner en plus des approximations des erreurs commises en d).

5. Toujours à partir des données de l'exercice 3:

a) Obtenir le système linéaire nécessaire pour calculer la spline naturelle dans l'intervalle $[0, 4]$.

b) Résoudre ce système et obtenir la valeur des dérivées secondes de la spline en chaque point d'interpolation.

c) Obtenir une approximation de $f(1{,}5)$ à l'aide de la spline.

6. Obtenir une approximation de $f(4{,}5)$ en utilisant un polynôme de degré 2 ainsi que les données suivantes.

x	$f(x)$
1,0	0,0000
2,0	0,6931
3,5	1,2528
5,0	1,6094
7,0	1,9459

a) Utiliser la méthode de Newton et un polynôme de degré 2. Donner l'expression analytique du terme d'erreur.

b) Répondre à la question posée en a), mais en utilisant cette fois la méthode de Lagrange.

c) Obtenir une approximation de l'erreur commise en a).

d) Est-ce possible d'obtenir une approximation de l'erreur commise en b)?

e) Quelles différences présentent ces deux méthodes?

7. Soit une fonction aléatoire X suivant une loi normale. La probabilité que X soit inférieur ou égal à x (notée $P(X \leq x)$) est donnée par la fonction:

$$P(X \leq x)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

Comme la fonction $e^{-t^2/2}$ n'a pas de primitive, on calcule cette probabilité pour différentes valeurs de x (par des méthodes que nous verrons au prochain chapitre) et on garde les résultats dans des tables comme celle-ci:

x	$P(X \leq x)$
1,0	0,841 34
1,1	0,864 33
1,2	0,884 93
1,3	0,903 20
1,4	0,919 24

À l'aide de cette table, obtenir $P(X \leq 1{,}05)$ avec une erreur absolue inférieure à $0{,}2 \times 10^{-5}$.

8. Un cas particulier intéressant de la formule d'interpolation de Newton se présente lorsque les points d'interpolation x_i sont également distants, c'est-à-dire lorsque:

$$x_{i+1} - x_i = h$$

Obtenir l'expression des premières, deuxièmes et troisièmes différences divisées dans ce cas précis. Donner un aperçu de ce que pourraient être les autres différences divisées.

9. Soit les trois points $(0\,,\,0)$, $(1\,,\,1)$ et $(2\,,\,8)$ de la fonction $f(x) = x^3$.

   a) Obtenir le système linéaire (1×1) permettant de calculer la spline cubique naturelle passant par ces trois points.

   b) À l'aide de la spline trouvée en a), donner une approximation de $f(1/2)$ et comparer le résultat avec la valeur exacte $1/8$.

   c) En interpolant une fonction cubique ($f(x) = x^3$) par des polynômes de degré 3 dans chaque intervalle, on obtient quand même une erreur. Expliquer.

10. Reprendre l'exercice précédent, mais en utilisant cette fois une spline qui vérifie $f_0'' = 0$ et $f_2'' = 12$. Expliquer les résultats.

11. On souhaite concevoir un virage d'une voie de chemin de fer entre les points $(0\,,\,0)$ et $(1\,,\,1)$. Le virage est décrit par une courbe de la forme $y = f(x)$ qui satisfait:

$$f(0) = 0 \quad \text{et} \quad f(1) = 1$$

De plus, pour assurer une transition en douceur, la pente de la courbe doit satisfaire:

$$f'(0) = 0 \quad \text{et} \quad f'(1) = 0{,}3$$

On représente la courbe à l'aide d'un polynôme dans l'intervalle $[0, 1]$.

a) Quel est le degré minimal que ce polynôme devra avoir pour remplir toutes les conditions?

b) Calculer ce polynôme.

12. Soit une fonction $f(x)$ dont on connaît la valeur en certains points.

x	$f(x)$
0	3
1	2
2	3
3	6
4	11
5	18

a) Calculer la table de différences divisées. Montrer que les troisièmes différences divisées sont nulles.

b) Que conclure au sujet de la fonction $f(x)$?

13. Soit les points de la table suivante.

x	$f(x)$
0	1
2	4
5	7

a) Construire le système linéaire de krigeage en utilisant la fonction $g(h) = h$ et une dérive linéaire.

b) Résoudre ce système par décomposition LU et donner l'expression de la fonction de krigeage $u(x)$. Évaluer $u(3)$.

c) Montrer qu'il s'agit bien d'une interpolation linéaire par morceaux.

14. Répondre aux questions a) et b) de l'exercice précédent, mais en utilisant cette fois la fonction $g(h) = h^3$.

15. En utilisant les mêmes points que ceux des deux exercices précédents, montrer que le système linéaire de krigeage obtenu en prenant la fonction $g(h) = h^2$ est singulier (ce qui indique que le choix de $g(h)$ n'est pas arbitraire).

16. Obtenir le système linéaire de krigeage qui permette de construire une surface passant par les points suivants.

x_1	x_2	$f(x_1, x_2)$
1	1	1
2	1	2
1	2	2
2	2	4

Évaluer la fonction de krigeage en $(3/2, 3/2)$ et comparer le résultat avec la valeur exacte $9/4$ (les points donnés appartiennent à la fonction $f(x_1, x_2) = x_1 x_2$). Utiliser pour ce faire une dérive linéaire et la fonction $g(h) = h^2 \ln h$.

17. À l'aide de la méthode du krigeage, donner l'équation paramétrique du carré $[0, 1] \times [0, 1]$.

    Suggestion: Considérer les 5 points $(0, 0)$, $(1, 0)$, $(1, 1)$, $(0, 1)$ et $(0, 0)$ de même que la fonction $g(h) = h$ pour obtenir une approximation linéaire par morceaux.

Chapitre 6

Différentiation et intégration numériques

6.1 Introduction

Le contenu de ce chapitre prolonge celui du chapitre 5 sur l'interpolation. À peu de choses près, on y manie les mêmes outils d'analyse. Dans le cas de l'interpolation, on cherchait à évaluer une fonction $f(x)$ connue seulement en quelques points. Dans le présent chapitre, le problème consiste à obtenir des approximations des différentes dérivées de cette fonction de même que de:

$$\int_{x_0}^{x_n} f(x)dx$$

On parle alors de *dérivation numérique* et d'*intégration numérique*. On fait face à ce type de problèmes lorsque, par exemple, on connaît la position d'une particule à intervalles de temps réguliers et que l'on souhaite obtenir sa vitesse. On doit alors effectuer la dérivée de la position connue seulement en quelques points. De même, l'accélération de cette particule nécessite le calcul de la dérivée seconde.

Si, à l'inverse, on connaît la vitesse d'une particule à certains intervalles de temps, on obtient la distance parcourue en intégrant la vitesse dans l'intervalle $[x_0, x_n]$.

Nous avons vu au chapitre précédent que la fonction $f(x)$ peut être convenablement estimée à l'aide d'un polynôme de degré n avec une certaine erreur. En termes concis:

$$f(x) = p_n(x) + E_n(x) \tag{6.1}$$

où $E_n(x)$ est le terme d'erreur d'ordre $(n+1)$ donné par la relation 5.21. L'expression 6.1 est à la base des développements de ce chapitre.

6.2 Différentiation numérique

On peut aborder la différentiation numérique d'au moins deux façons. La première approche consiste à utiliser le développement de Taylor et la seconde est fondée sur l'égalité 6.1. Nous utiliserons un mélange des deux approches, ce qui nous permettra d'avoir un portrait assez complet de la situation.

Commençons d'abord par l'équation 6.1. Si on dérive de chaque côté de l'égalité, on obtient successivement:

$$\begin{array}{rcl} f'(x) & = & p_n'(x) + E_n'(x) \\ f''(x) & = & p_n''(x) + E_n''(x) \\ f'''(x) & = & p_n'''(x) + E_n'''(x) \\ \vdots & = & \vdots \end{array} \tag{6.2}$$

Ainsi, pour évaluer la dérivée d'une fonction connue aux points $((x_i, f(x_i))$ pour $i = 0, 1, 2, \cdots, n)$, il suffit de dériver le polynôme d'interpolation passant par ces points. De plus, *le terme d'erreur associé à cette approximation de la dérivée est tout simplement la dérivée de l'erreur d'interpolation.* Ce résultat est vrai quel que soit l'ordre de la dérivée.

Remarque 6.1
Bien qu'en théorie on soit en mesure d'estimer les dérivées de tout ordre, sur le plan pratique, on dépasse rarement l'ordre 4. Cela s'explique par le fait que la différentiation numérique est un procédé numériquement instable. □

6.2.1 Dérivées d'ordre 1

Commençons par faire l'approximation des dérivées d'ordre 1, ce qui revient à évaluer la pente de la fonction $f(x)$. Tout comme pour l'interpolation, nous avons le choix entre plusieurs polynômes de degré plus ou moins élevé. De ce choix dépendent l'ordre et la précision de l'approximation. Nous avons rencontré un problème semblable dans le cas de l'interpolation: si un polynôme de degré n est utilisé, on obtient une approximation d'ordre $(n+1)$ de la fonction $f(x)$ (voir la relation 5.25).

Il est également utile de se rappeler que l'erreur d'interpolation s'écrit:

$$E_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}[(x-x_0)(x-x_1)\cdots(x-x_n)] \tag{6.3}$$

pour un certain ξ compris dans l'intervalle $[x_0, x_n]$. En dérivant cette expression, tout en tenant compte de la dépendance de ξ envers x, on obtient:

$$\begin{aligned} E_n'(x) &= \frac{f^{(n+2)}(\xi(x))\xi'(x)}{(n+1)!}[(x-x_0)(x-x_1)\cdots(x-x_n)] \\ &+ \frac{f^{(n+1)}(\xi(x))}{(n+1)!}[(x-x_0)(x-x_1)\cdots(x-x_n)]' \end{aligned}$$

La dérivée du produit apparaissant dans le deuxième terme de droite est plus délicate. Cette dérivée débouche sur une somme de produits où, tour à tour, l'un des facteurs $(x - x_i)$ est manquant. Il est facile de se convaincre, en reprenant ce développement avec $n = 2$ par exemple, que l'on obtient:

$$\begin{aligned} E_n'(x) &= \frac{f^{(n+2)}(\xi(x))\xi'(x)}{(n+1)!}[(x-x_0)(x-x_1)\cdots(x-x_n)] \\ &+ \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\left(\sum_{k=0}^{n}\prod_{j=0(j\neq k)}^{n}(x-x_j)\right) \end{aligned} \tag{6.4}$$

On peut simplifier cette expression quelque peu complexe en choisissant l'uu ou l'autre des points d'interpolation. En effet, en $x = x_i$, le premier terme de droite s'annule, faisant disparaître la dérivée de $\xi(x)$, qui est inconnue. De la somme, il ne reste qu'un seul terme puisque tous les autres contiennent un facteur $(x - x_i)$ et s'annulent. Il reste:

$$E_n'(x_i) = \frac{f^{(n+1)}(\xi(x_i))}{(n+1)!}\left(\prod_{j=0(j\neq i)}^{n}(x_i-x_j)\right)$$

Si on suppose de plus que les x_i sont également distancés, c'est-à-dire:

$$x_{i+1} - x_i = h$$

ce qui signifie que $x_i - x_j = (i-j)h$, on obtient:

$$E_n'(x_i) = \frac{f^{(n+1)}(\xi_i)h^n}{(n+1)!}\left(\prod_{j=0(j\neq i)}^{n}(i-j)\right) \tag{6.5}$$

où ξ_i est simplement une notation différente de $\xi(x_i)$. En particulier, si $i = 0$, *on trouve:*

$$E_n'(x_0) = \frac{f^{(n+1)}(\xi_0)h^n}{(n+1)!}\left(\prod_{j=0(j\neq 0)}^{n}(-j)\right) = \frac{f^{(n+1)}(\xi_0)h^n}{(n+1)!}\left(\prod_{j=1}^{n}(-j)\right)$$

c'est-à-dire:

$$E_n'(x_0) = \frac{(-1)^n h^n f^{(n+1)}(\xi_0)}{(n+1)} \qquad (6.6)$$

pour un certain ξ_0 compris dans l'intervalle $[x_0, x_n]$.

Remarque 6.2
L'équation 6.5 montre que si on utilise un polynôme d'interpolation de degré n (c'est-à-dire d'ordre $(n+1)$) la dérivée de ce polynôme évaluée en $x = x_i$ est une approximation d'ordre n de $f'(x_i)$. □

Définition 6.1
Aux points d'interpolation, on a:

$$f'(x_i) = p_n'(x_i) + E_n'(x_i) \qquad (6.7)$$

Le terme $p_n'(x_i)$ dans l'équation 6.7 est une *formule aux différences finies* ou plus simplement une formule aux différences. Nous proposons plus loin plusieurs formules aux différences finies pour évaluer les différentes dérivées de $f(x)$. Elles se distinguent principalement par le degré du polynôme et par les points d'interpolation retenus.

Exemple 6.1
Si on choisit le polynôme de degré 1 passant par les points $(x_0, f(x_0))$ et $(x_1, f(x_1))$, on a, grâce à la formule d'interpolation de Newton:

$$p_1(x) = f(x_0) + f[x_0, x_1](x - x_0)$$

et donc:

$$f'(x) = p_1'(x) + E_1'(x) = f[x_0, x_1] + E_1'(x) \qquad (6.8)$$

En vertu de la relation 6.6 avec $n = 1$ et puisque $(x_1 - x_0) = h$, on arrive à:

$$f'(x_0) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} + E_1'(x_0) = \frac{f(x_1) - f(x_0)}{h} + \frac{(-1)^1 h^1 f^{(2)}(\xi_0)}{2}$$

qui peut encore s'écrire:

$$f'(x_0) = \frac{f(x_1) - f(x_0)}{h} - \frac{hf^{(2)}(\xi_0)}{2} \quad \text{pour } \xi_0 \in [x_0, x_1] \tag{6.9}$$

qui est la *différence avant d'ordre 1.* On l'appelle différence avant car, pour évaluer la dérivée en $x = x_0$, on cherche de l'information vers l'avant (en $x = x_1$).
De la même manière, si on évalue l'équation 6.8 en $x = x_1$, la relation 6.5 avec $(i = 1)$ donne:

$$\begin{aligned} f'(x_1) &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} + E_1'(x_1) \\ &= \frac{f(x_1) - f(x_0)}{h} + \frac{h^1 f^{(2)}(\xi_1)}{2!} \left(\prod_{j=0 (j \neq 1)}^{1} (1 - j) \right) \end{aligned}$$

ou encore:

$$f'(x_1) = \frac{f(x_1) - f(x_0)}{h} + \frac{hf^{(2)}(\xi_1)}{2} \quad \text{pour } \xi_1 \in [x_0, x_1] \tag{6.10}$$

qui est la *différence arrière d'ordre 1.*

● ● ● ● ●

Remarque 6.3
L'exemple précédent montre que la même différence divisée est une approximation de la dérivée à la fois en $x = x_0$ et en $x = x_1$. On remarque cependant que le terme d'erreur est différent aux deux endroits. □

Exemple 6.2
Passons maintenant aux polynômes de degré 2. Soit les points $(x_0, f(x_0))$, $(x_1, f(x_1))$ et $(x_2, f(x_2))$. Le polynôme de degré 2 passant par ces trois points est:

$$p_2(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)$$

dont la dérivée est:

$$p_2'(x) = f[x_0, x_1] + f[x_0, x_1, x_2](2x - (x_0 + x_1))$$

Lorsque x prend successivement les valeurs x_0, x_1 et x_2, il est facile de montrer que l'on obtient des approximations d'ordre 2 de la dérivée.

$f'(x_0) = \dfrac{-f(x_2) + 4f(x_1) - 3f(x_0)}{2h} + \dfrac{h^2 f'''(\xi_0)}{3}$ *Différence avant d'ordre 2*
$f'(x_1) = \dfrac{f(x_2) - f(x_0)}{2h} - \dfrac{h^2 f'''(\xi_1)}{6}$ *Différence centrée d'ordre 2*
$f'(x_2) = \dfrac{3f(x_2) - 4f(x_1) + f(x_0)}{2h} + \dfrac{h^2 f'''(\xi_2)}{3}$ *Différence arrière d'ordre 2*

● ● ● ● ●

Remarque 6.4
Les termes d'erreur de ces formules aux différences finies découlent tous de la relation 6.5 lorsqu'on pose successivement $i = 0, 1$ et 2. Pour $i = 0$, on peut utiliser directement l'équation 6.6. Les points ξ_0, ξ_1 et ξ_2 sont situés quelque part dans l'intervalle $[x_0, x_2]$ et sont inconnus (voir les exercices de fin de chapitre). □

Remarque 6.5
Toutes ces formules aux différences sont d'ordre 2. Les mentions avant, centrée et arrière renvoient au point où on calcule la dérivée et aux points utilisés pour la calculer. Ainsi, la différence avant est évaluée en x_0 sur la base des valeurs situées vers l'avant, soit en x_1 et en x_2. La différence arrière fixe la dérivée en $x = x_2$ avec l'appui des valeurs de la fonction en x_0 et en x_1. La différence centrée, quant à elle, fait intervenir des valeurs situées de part et d'autre de x_1.

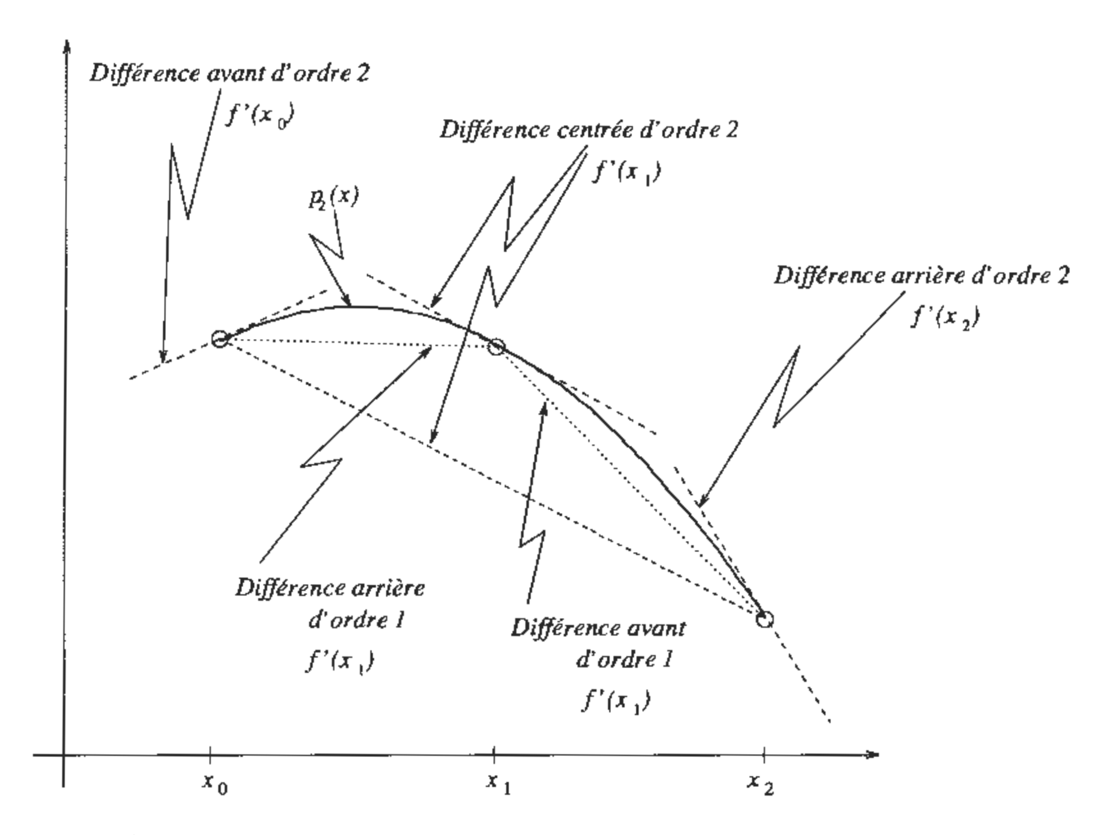


Figure 6.1: Interprétation géométrique des formules aux différences

La figure 6.1 illustre les différentes possibilités. Pour les différences d'ordre 1, on estime la dérivée par la pente du segment de droite joignant les points $(x_0, f(x_0))$ et $(x_1, f(x_1))$. Dans le cas des différences d'ordre 2, on détermine un polynôme de degré 2 dont la pente en x_0, en x_1 et en x_2 donne respectivement les différences avant, centrée et arrière. □

On peut aussi convenir de toujours évaluer la dérivée en x. Dans ce cas, on utilise les valeurs de $f(x + h)$ et de $f(x + 2h)$ pour la différence avant et les valeurs de $f(x + h)$ et de $f(x - h)$ pour la différence centrée. En ce qui concerne le terme d'erreur, on ne retient que son ordre. Le tableau suivant résume la situation.

$$f'(x) = \frac{f(x+h) - f(x)}{h} + O(h)$$

Différence avant d'ordre 1

$$f'(x) = \frac{f(x) - f(x-h)}{h} + O(h)$$

Différence arrière d'ordre 1

$$f'(x) = \frac{-f(x+2h) + 4f(x+h) - 3f(x)}{2h} + O(h^2) \qquad (6.11)$$

Différence avant d'ordre 2

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} + O(h^2)$$

Différence centrée d'ordre 2

$$f'(x) = \frac{3f(x) - 4f(x-h) + f(x-2h)}{2h} + O(h^2)$$

Différence arrière d'ordre 2

Exemple 6.3
On tente d'évaluer la dérivée de $f(x) = e^x$ en $x = 0$. La solution exacte est dans ce cas $f'(0) = e^0 = 1$. On peut dès lors comparer ce résultat avec ceux que l'on obtient par les différentes formules aux différences. Par exemple, la différence avant d'ordre 1 donne pour $h = 0{,}1$:

$$f'(0) \simeq \frac{e^{0+h} - e^0}{h} = \frac{e^{0{,}1} - 1}{0{,}1} = 1{,}051\,709\,18$$

Une valeur plus petite de h conduit à un résultat plus précis. Ainsi, si $h = 0{,}05$:

$$f'(0) \simeq \frac{e^{0,05} - 1}{0{,}05} = 1{,}025\,4219$$

On obtient ainsi une erreur à peu près deux fois plus petite, ce qui confirme que cette approximation est d'ordre 1. Si on utilise cette fois une différence centrée d'ordre 2, on obtient avec $h = 0{,}05$:

$$f'(0) \simeq \frac{e^{0,05} - e^{-0,05}}{2(0{,}05)} = 1{,}000\,4167$$

qui est un résultat beaucoup plus précis. Avec $h = 0{,}025$, on obtient:

$$f'(0) \simeq \frac{e^{0,025} - e^{-0,025}}{2(0{,}025)} = 1{,}000\,104\,18$$

soit une erreur à peu près 4 fois plus petite qu'avec $h = 0{,}05$. On obtiendrait des résultats similaires avec les différences avant et arrière d'ordre 2.

• • • • •

6.2.2 Dérivées d'ordre supérieur

Avec les dérivées d'ordre supérieur, on agit à peu près de la même manière qu'avec les dérivées d'ordre 1, c'est-à-dire que l'on dérive un polynôme d'interpolation aussi souvent que nécessaire. Les dérivées d'ordre supérieur posent toutefois une difficulté supplémentaire, qui provient principalement de l'analyse d'erreur. En effet, dériver plusieurs fois le terme d'erreur 5.21 est long et fastidieux. Nous préférons suivre une approche légèrement différente basée sur le développement de Taylor.

Reprenons le polynôme de degré 2 déjà utilisé pour calculer la dérivée première. Ce polynôme s'écrit:

$$p_2(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)$$

et sa dérivée seconde est:

$$p_2''(x) = 2f[x_0, x_1, x_2] = \frac{f(x_2) - 2f(x_1) + f(x_0)}{h^2} \tag{6.12}$$

qui constitue une approximation de la dérivée seconde $f''(x)$ partout dans l'intervalle $[x_0, x_2]$. Il reste à en déterminer l'ordre. Cet ordre dépend du point retenu pour l'approximation.

- Premier cas: On fait l'approximation de la dérivée en x_0.

  L'équation 6.12 peut alors s'écrire:

$$f''(x_0) \simeq p_2''(x_0) = \frac{f(x_0 + 2h) - 2f(x_0 + h) + f(x_0)}{h^2}$$

  On remarque immédiatement qu'il s'agit d'une formule aux différences avant. Ponr déterminer l'ordre de l'erreur liée à cette approximation, on utilise le développement de Taylor 1.24. Dans un premier temps, on a:

$$\begin{aligned} f(x_0 + 2h) &= f(x_0) + f'(x_0)(2h) + \frac{f''(x_0)}{2!}(2h)^2 \\ &+ \frac{f'''(x_0)}{3!}(2h)^3 + \frac{f''''(x_0)}{4!}(2h)^4 + \cdots \end{aligned}$$

  et de même:

$$\begin{aligned} f(x_0 + h) &= f(x_0) + f'(x_0)h + \frac{f''(x_0)}{2!}h^2 \\ &+ \frac{f'''(x_0)}{3!}h^3 + \frac{f''''(x_0)}{4!}h^4 + \cdots \end{aligned}$$

  On parvient alors à:

$$\begin{aligned} \frac{f(x_0 + 2h) - 2f(x_0 + h) + f(x_0)}{h^2} &= \frac{f''(x_0)h^2 + f'''(x_0)h^3 + O(h^4)}{h^2} \\ &= f''(x_0) + f'''(x_0)h + O(h^2) \\ &= f''(x_0) + O(h) \end{aligned}$$

  Cette différence avant est donc une approximation d'ordre 1 de la dérivée seconde. *C'est cette approximation que l'on a utilisée pour évaluer l'erreur d'interpolation 5.21 à l'aide de la formule 5.22 au chapitre précédent.*

- Deuxième cas: On fait l'approximation de la dérivée en x_1.

  L'éqnation 6.12 peut alors s'écrire:

$$f''(x_1) \simeq p_2''(x_1) = \frac{f(x_1 + h) - 2f(x_1) + f(x_1 - h)}{h^2}$$

qui est une différence centrée. Pour en déterminer l'ordre, on fait appel, comme dans le cas précédent, aux développements de Taylor, mais cette fois autour de x_1. On a:

$$\begin{aligned} f(x_1+h) &= f(x_1) + f'(x_1)h + \frac{f''(x_1)}{2!}h^2 \\ &+ \frac{f'''(x_1)}{3!}h^3 + \frac{f''''(x_1)}{4!}h^4 + \cdots \end{aligned}$$

En remplaçant h par $(-h)$, on obtient également:

$$\begin{aligned} f(x_1-h) &= f(x_1) - f'(x_1)h + \frac{f''(x_1)}{2!}h^2 \\ &- \frac{f'''(x_1)}{3!}h^3 + \frac{f''''(x_1)}{4!}h^4 + \cdots \end{aligned}$$

Une fois combinées, ces deux relations deviennent:

$$\begin{aligned} \frac{f(x_1+h) - 2f(x_1) + f(x_1-h)}{h^2} &= \frac{f''(x_1)h^2 + \frac{f''''(x_1)}{12}h^4 + O(h^6)}{h^2} \\ &= f''(x_1) + \frac{f''''(x_1)}{12}h^2 + O(h^4) \\ &= f''(x_1) + O(h^2) \end{aligned}$$

c'est-à-dire une approximation d'ordre 2 de la dérivée.

- Troisième cas: On fait l'approximation de la dérivée en x_2.

  En reprenant un raisonnement similaire à celui dn premier cas, on pourrait montrer que la relation 6.12 est une approximation d'ordre 1 de la dérivée seconde en $x = x_2$.

Remarque 6.6
Il pent sembler surprenant de constater que la même équation aux différences, obtenue à partir d'un polynôme de degré 2, soit d'ordre 1 en $x = x_0$ et en $x = x_2$ et soit d'ordre 2 en $x = x_1$. Cela s'explique par la symétrie des différences centrées, qui permet de gagner un ordre de précision. □

On peut obtenir toute une série de formules aux différences finies en utilisant des polynômes de degré plus ou moins élevé et en choisissant les développements de Taylor appropriés pour en obtenir l'ordre de convergence. Le tableau suivant présente les principales d'entres elles.

$$f''(x) = \frac{f(x-2h) - 2f(x-h) + f(x)}{h^2} + O(h)$$

Différence arrière d'ordre 1

$$f''(x) = \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2} + O(h)$$

Différence avant d'ordre 1

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} + O(h^2) \tag{6.13}$$

Différence centrée d'ordre 2

$$f''(x) = \frac{-f(x+2h)+16f(x+h)-30f(x)+16f(x-h)-f(x-2h)}{12h^2} + O(h^4)$$

Différence centrée d'ordre 4

$$f''''(x) = \frac{f(x+2h)-4f(x+h)+6f(x)-4f(x-h)+f(x-2h)}{h^4} + O(h^4)$$

Différence centrée d'ordre 4

Pour terminer, nous démontrons que la différentiation est un procédé numériquement instable. Toutes les formules de différences finies dépendent d'un paramètre h qui est la distance entre les points d'interpolation. On pourrait croire, de façon intuitive, que la précision du résnltat augmente à mesure que diminue la valeur de h. Dans le cas de la différentiation numérique, il y a une limite aux valeurs de h qui peuvent être utilisées. En effet, si on prend, par exemple, une différence centrée pour estimer la dérivée première, c'est-à-dire:

$$f'(x_0) \simeq \frac{f(x_0+h) - f(x_0-h)}{2h}$$

on constate que lorsque h tend vers 0 le numérateur contient la soustraction de deux termes très proches l'un de l'autre. Cela résulte en l'élimination par

soustraction (voir la section 1.5.2) de plusieurs chiffres significatifs lorsque h est trop petit. À quoi s'ajoute une division par un nombre très petit. L'exemple suivant illustre ce phénomène.

Exemple 6.4

On considère les différences centrées d'ordre 2 pour le calcul des dérivées première et deuxième de la fonction $f(x) = e^x$ en $x = 0$. Ces deux calculs, qui doivent normalement aboutir à 1, permettent d'apprécier la précision des résultats. Le tableau suivant rassemble les résultats en simple précision (IEEE), ce qui correspond à peu près à travailler avec une mantisse de 7 chiffres décimaux.

h	$f'(x) \simeq \frac{f(x+h)-f(x-h)}{2h}$	$f''(x) \simeq \frac{f(x+h)-2f(x)+f(x-h)}{h^2}$
$0{,}1 \times 10^{+1}$	1,175 201 178	1,086 161 137
$0{,}1 \times 10^{+0}$	1,001 667 619	1,000 839 472
$0{,}1 \times 10^{-2}$	1,000 016 928	1,000 165 939
$0{,}1 \times 10^{-3}$	1,000 017 047	1,013 279 080
$0{,}1 \times 10^{-4}$	1,000 166 059	0,000 000 000
$0{,}1 \times 10^{-5}$	1,001 358 151	0,000 000 000
$0{,}1 \times 10^{-6}$	0,983 476 758	−59 604,660 16

La valeur de h est successivement réduite d'un facteur de 10 à partir de $h = 1{,}0$. On constate que lorsque h diminue la précision liée aux dérivées augmente dans un premier temps, puis se dégrade brusquement pour les valeurs de h plus faibles. Cela est particulièrement évident pour la dérivée seconde (troisième colonne). Si on passe en double précision (l'équivalent de 14 chiffres décimaux dans la mantisse), on observe un comportement

similaire, mais qui se produit à des valeurs de h plus faibles.

h	$f'(x) \simeq \frac{f(x+h)-f(x-h)}{2h}$	$f''(x) \simeq \frac{f(x+h)-2f(x)+f(x-h)}{h^2}$
$0{,}1 \times 10^{+01}$	1,175 201 193 643 801 38	1,086 161 269 630 487 42
$0{,}1 \times 10^{-00}$	1,001 667 500 198 440 97	1,000 833 611 160 722 78
$0{,}1 \times 10^{-01}$	1,000 016 666 749 992 12	1,000 008 333 360 558 05
$0{,}1 \times 10^{-02}$	1,000 000 166 666 681 34	1,000 000 083 406 504 81
$0{,}1 \times 10^{-03}$	1,000 000 001 666 889 74	1,000 000 005 024 759 28
$0{,}1 \times 10^{-04}$	1,000 000 000 012 102 32	0,999 998 972 517 346 26
$0{,}1 \times 10^{-05}$	0,999 999 999 973 244 40	0,999 977 878 279 878 16
$0{,}1 \times 10^{-06}$	0,999 999 999 473 643 93	0,999 200 722 162 640 44
$0{,}1 \times 10^{-07}$	0,999 999 993 922 528 80	0,000 000 000 000 000 00
$0{,}1 \times 10^{-08}$	1,000 000 027 229 219 55	111,022 302 462 515 597
$0{,}1 \times 10^{-09}$	1,000 000 082 740 370 78	0,000 000 000 000 000 00
$0{,}1 \times 10^{-10}$	1,000 000 082 740 370 78	0,000 000 000 000 000 00
$0{,}1 \times 10^{-11}$	1,000 033 389 431 109 53	111 022 302,462 515 622
$0{,}1 \times 10^{-12}$	0,999 755 833 674 953 35	−11 102 230 246,251 562 1
$0{,}1 \times 10^{-13}$	0,999 200 722 162 640 77	0,000 000 000 000 000 000
$0{,}1 \times 10^{-14}$	1,054 711 873 393 898 71	111 022 302 462 515,641
$0{,}1 \times 10^{-15}$	0,555 111 512 312 578 15	−11 102 230 246 251 564,0
$0{,}1 \times 10^{-16}$	0,000 000 000 000 000 00	0,000 000 000 000 000 00
$0{,}1 \times 10^{-17}$	0,000 000 000 000 000 00	0,000 000 000 000 000 00
$0{,}1 \times 10^{-18}$	0,000 000 000 000 000 00	0,000 000 000 000 000 00

Lorsque h est trop petit, l'élimination par soustraction des chiffres significatifs a un impact dévastateur sur la précision des résultats. Il est donc recommandé d'être très prudent dans le choix de h et d'éviter des valeurs trop petites.

● ● ● ● ●

6.3 Extrapolation de Richardson

La méthode d'extrapolation de Richardson est valable non seulement pour la différentiation et l'intégration numériques, mais aussi pour l'interpolation, la résolution numérique des équations différentielles, etc. Cette technique permet d'augmenter la précision d'une méthode d'approximation par une technique d'extrapolation que nous décrivons dans cette section.

Prenons comme point de départ une approximation numérique, notée $Q_{app}(h)$, d'une certaine quantité exacte Q_{exa} inconnue. L'approximation numérique dépend d'un paramètre h, comme c'est souvent le cas. Généralement, plus h est petit, plus l'approximation est précise. On suppose de plus que cette approximation est d'ordre n, c'est-à-dire:

$$Q_{exa} = Q_{app}(h) + O(h^n)$$

La notation $O(h^n)$ signifie en fait que l'on a:

$$Q_{exa} = Q_{app}(h) + c_n h^n + c_{n+1} h^{n+1} + c_{n+2} h^{n+2} + \cdots \qquad (6.14)$$

où les constantes c_n dépendent de la méthode nnmérique utilisée. La technique d'extrapolation de Richardson consiste à obtenir, à partir de l'approximation 6.14 d'ordre n, une nouvelle approximation d'ordre *au moins* $(n+1)$. Ponr ce faire, il suffit de remplacer h par $h/2$ dans l'équation 6.14, ce qui conduit à la relation:

$$Q_{exa} = Q_{app}(\frac{h}{2}) + c_n(\frac{h}{2})^n + c_{n+1}(\frac{h}{2})^{n+1} + c_{n+2}(\frac{h}{2})^{n+2} + \cdots \qquad (6.15)$$

L'approximation $Q_{app}(h/2)$ est généralement plus précise que $Q_{app}(h)$. On peut cependant se servir de ces deux approximations ponr en obtenir une nouvelle, encore plus précise. L'idée consiste à combiner les relations 6.14 et 6.15 de telle sorte que le terme d'ordre n ($c_n h^n$) disparaisse. Cela est possible si on multiplie l'équation 6.15 par 2^n pour obtenir:

$$2^n Q_{exa} = 2^n Q_{app}(\frac{h}{2}) + c_n h^n + c_{n+1}(\frac{h^{n+1}}{2}) + c_{n+2}(\frac{h^{n+2}}{2^2}) + \cdots$$

En sonstrayant l'énoncé 6.14 de cette dernière relation, on obtient:

$$(2^n - 1)Q_{exa} = 2^n Q_{app}(\frac{h}{2}) - Q_{app}(h) - \frac{1}{2}c_{n+1}h^{n+1} - \frac{3}{4}c_{n+2}h^{n+2} + \cdots$$

d'où:

$$Q_{exa} = \frac{2^n Q_{app}(\frac{h}{2}) - Q_{app}(h)}{(2^n - 1)} + \frac{-\frac{1}{2}c_{n+1}h^{n+1} - \frac{3}{4}c_{n+2}h^{n+2} + \cdots}{(2^n - 1)} \qquad (6.16)$$

qui s'écrit plus simplement:

$$Q_{exa} = \frac{2^n Q_{app}(\frac{h}{2}) - Q_{app}(h)}{(2^n - 1)} + O(h^{n+1})$$

L'expression de droite est donc une approximation d'ordre au moins $(n+1)$ de Q_{exa}. *L'extrapolation de Richardson permet donc de gagner au moins un ordre de convergence.* En fait, on peut eu gagner davantage si, par exemple, on a $c_{n+1} = 0$ dès le départ. Dans ce cas, la nouvelle approximation est d'ordre $(n+2)$. Cette situation se produit fréquemment, notamment avec les différences centrées et la méthode d'intégration dite des trapèzes que nous verrons plus loin.

Exemple 6.5
On a vu qu'en utilisant une différence avant d'*ordre 1* pour calculer la dérivée de e^x en $x = 0$ on obtient:

$$f'(0) \simeq \frac{e^{0+h} - e^0}{h} = \frac{e^{0,1} - 1}{0,1} = 1,051\,709\,18 = Q_{app}(0,1)$$

pour $h = 0,1$ et:

$$f'(0) \simeq \frac{e^{0,05} - 1}{0,05} = 1,025\,4219 = Q_{app}(0,05)$$

pour $h = 0,05$. On peut maintenant faire le calcul à l'aide de l'équation 6.16 avec $n = 1$:

$$\begin{aligned} f'(0) &\simeq \frac{2^1 Q_{app}(0,05) - Q_{app}(0,1)}{2^1 - 1} \\ &= (2)(1,025\,421\,9) - 1,051\,709\,18 = 0,999\,134\,62 \end{aligned}$$

qui est une approximation d'ordre 2 et donc plus précise de $f'(0)$. De même, si on utilise une différence centrée d'ordre 2, on obtient pour $h = 0,05$:

$$f'(0) \simeq \frac{e^{0,05} - e^{-0,05}}{2(0,05)} = 1,000\,4167$$

et avec $h = 0,025$:

$$f'(0) \simeq \frac{e^{0,025} - e^{-0,025}}{2(0,025)} = 1,000\,104\,18$$

Dans ce cas, l'extrapolation de Richardson permet de gagner 2 ordres de précision puisque seules les puissances paires de h apparaissent dans le terme

d'erreur (voir les exercices de fin de chapitre). Plus précisément, on a:

$$\frac{f(x+h)-f(x-h)}{2h} = f'(x) + \frac{f'''(x)h^2}{3!} + \frac{f^{(5)}(x)h^4}{5!} + O(h^6)$$

La différence centrée étant d'ordre 2, l'extrapolation de Richardson avec $n = 2$ donne:

$$\begin{aligned} f'(0) &\simeq \frac{2^2 Q_{app}(0{,}025) - Q_{app}(0{,}05)}{2^2 - 1} \\ &= \frac{(4)(1{,}000\,104\,18) - 1{,}000\,4167}{3} = 1{,}000\,000\,007 \end{aligned}$$

qui est une approximation d'ordre 4 de la solution exacte.

● ● ● ● ●

Remarque 6.7
Des exemples précédents, on conclut qu'il vaut mieux éviter d'utiliser des valeurs de h très petites pour calculer une dérivée à l'aide d'une formule de différences finies. Il est en effet préférable de choisir une valeur de h pas trop petite et de faire des extrapolations de Richardsou. □

6.4 Intégration numérique

L'intégration numérique est basée principalement sur la relation:

$$\int_{x_0}^{x_n} f(x)dx = \int_{x_0}^{x_n} p_n(x)dx + \int_{x_0}^{x_n} E_n(x)dx \qquad (6.17)$$

où $p_n(x)$ est un polynôme d'interpolatiou et $E_n(x)$ est l'erreur qui y est associée. En faisant varier la valeur de n, on obtient les *formules de Newton-Cotes*. En principe, plus n est élevé et plus grande est la précision liée à la valeur de l'intégrale recherchée. En pratique cependant, les numériciens emploient des valeurs de n inférieures ou égales à 5.

Par ailleurs, l'extrapolation de Richardson, alliée judicieusement à l'une des formules de Newton-Cotes, conduit à la méthode de Romberg, l'une des techniques d'intégration numérique les plus précises. Enfin, nous traitons des quadratures de Gauss, très fréquemment utilisées dans les méthodes numériques plus avancées comme celle des éléments finis (voir Reddy, réf. [24]).

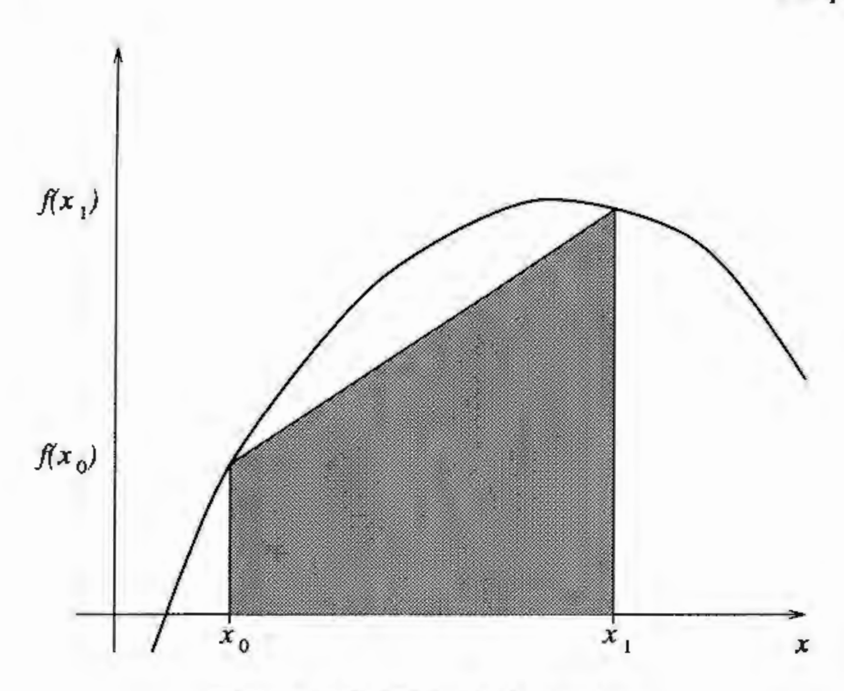


Figure 6.2: Méthode du trapèze

6.4.1 Formules de Newton-Cotes simples et composées

Méthode des trapèzes

Commençons par la méthode la plus simple. On souhaite évaluer:

$$\int_{x_0}^{x_1} f(x)dx$$

où $f(x)$ est une fonction connue seulement en deux points ou encore une fonction n'ayant pas de primitive. La solution qui vient tout de suite à l'esprit consiste à remplacer $f(x)$ par le polynôme de degré 1 passant par les points $(x_0, f(x_0))$ et $(x_1, f(x_1))$ comme l'illustre la figure 6.2.

La valeur approximative de l'intégrale correspond à l'aire sous la courbe du polynôme. Cette aire forme un trapèze qui donne son nom à la *méthode du trapèze*. Évidemment, l'approximation est grossière et on peut d'ores et déjà soupçonner que le résultat sera peu précis. Le polynôme de Newton 5.6

et la relation 5.21 conduisent à:

$$\begin{aligned}\int_{x_0}^{x_1} f(x)dx &= \int_{x_0}^{x_1} p_1(x)dx + \int_{x_0}^{x_1} E_1(x)dx \\ &= \int_{x_0}^{x_1} \{f(x_0) + f[x_0, x_1](x - x_0)\}dx \\ &\quad + \int_{x_0}^{x_1} \frac{f''(\xi(x))}{2!}(x - x_0)(x - x_1)dx\end{aligned}$$

ce qui peut également s'écrire, si on intègre le polynôme:

$$\begin{aligned}\int_{x_0}^{x_1} f(x)dx &= \frac{(x_1 - x_0)}{2}(f(x_0) + f(x_1)) \\ &\quad + \int_{x_0}^{x_1} \frac{f''(\xi(x))}{2!}(x - x_0)(x - x_1)dx\end{aligned} \tag{6.18}$$

Le premier terme de droite n'est rien d'autre que l'aire du trapèze de la figure 6.2, tandis que le deuxième terme est l'erreur commise. Le changement de variable 5.24 permet d'écrire:

$$s = \frac{x - x_0}{h}$$

d'où l'on tire que $(x - x_i) = (s - i)h$ et que $dx = hds$. Le terme d'erreur devient alors:

$$\int_{x_0}^{x_1} \frac{f''(\xi(x))}{2!}(x - x_0)(x - x_1)dx = \int_0^1 \frac{f''(\xi(s))}{2!}s(s - 1)h^3 ds$$

On peut encore simplifier cette expression en faisant appel au second théorème de la moyenne.

Théorème 6.1

Soit $f_1(x)$, une fonction continue dans l'intervalle $[a, b]$ et $f_2(x)$, une fonction intégrable qui ne change pas de signe dans l'intervalle $[a, b]$. Il existe alors $\eta \in [a, b]$ tel que:

$$\int_a^b f_1(x)f_2(x)dx = f_1(\eta)\int_a^b f_2(x)dx \qquad \square \tag{6.19}$$

Comme la fonction $(s(s-1))$ ne change pas de signe dans $[0,1]$, on peut mettre à profit ce théorème, ce qui donne:

$$\int_0^1 \frac{f''(\xi(s))}{2!} s(s-1)h^3 ds = \frac{f''(\eta)}{2!}h^3 \int_0^1 s(s-1)ds = -\frac{f''(\eta)}{12}h^3$$

La méthode du trapèze se résume donc à l'égalité:

$$\int_{x_0}^{x_1} f(x)dx = \frac{h}{2}(f(x_0) + f(x_1)) - \frac{f''(\eta)}{12}h^3 \quad \text{pour} \quad \eta \in [x_0, x_1] \qquad (6.20)$$

La méthode du trapèze demeure peu précise, comme en témoigne l'exemple suivant.

Exemple 6.6
Il s'agit d'évaluer numériquement:

$$\int_0^{\frac{\pi}{2}} \sin x dx$$

dont la valeur exacte est 1. La méthode du trapèze donne dans ce cas:

$$\int_0^{\frac{\pi}{2}} \sin x dx \simeq \frac{\frac{\pi}{2}}{2}(\sin 0 + \sin \frac{\pi}{2}) = \frac{\pi}{4} = 0{,}785\,398\,164$$

qui est une piètre approximation de la valeur exacte 1.

● ● ● ● ●

Ce résultat peu impressionnant vient du fait que l'on approche la fonction $\sin x$ dans l'intervalle $[0, \frac{\pi}{2}]$ au moyen d'un polynôme de degré 1. Cette approximation est assez médiocre, comme en témoigne la figure 6.3.

Une stratégie intéressante consiste à décomposer l'intervalle où l'on doit faire l'intégration, soit l'intervalle $[a,b]$, en n sous-intervalles de longueur (voir la figure 6.4):

$$h = \frac{b-a}{n} \qquad (6.21)$$

Les différents points engendrés sont notés x_i pour $i = 0, 1, 2, \cdots, n$. Les valeurs aux extrémités sont $a = x_0$ et $b = x_n$. Dans chaque sous-intervalle $[x_i, x_{i+1}]$, on peut utiliser la méthode du trapèze. On a alors:

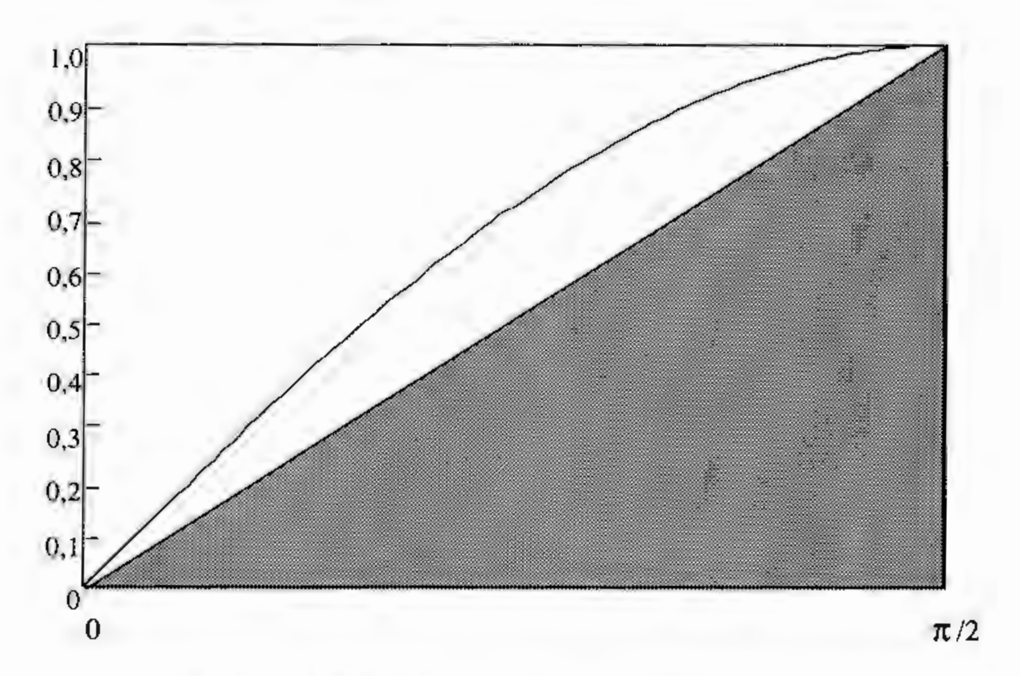


Figure 6.3: Méthode du trapèze, $f(x) = \sin x$

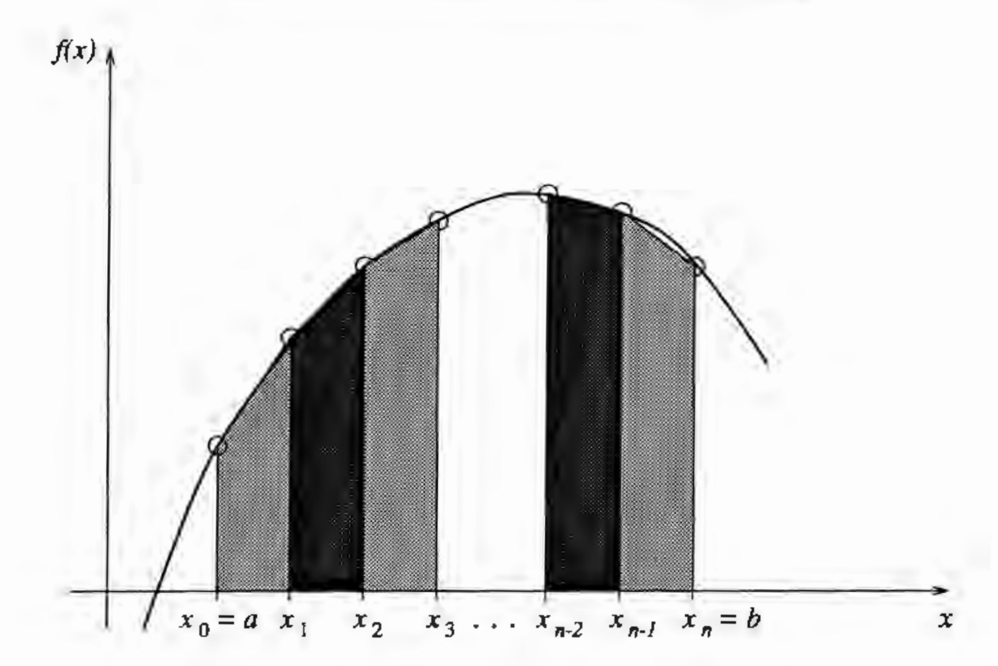


Figure 6.4: Méthode des trapèzes composée

$$\begin{aligned}\int_a^b f(x)dx &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)dx \simeq \sum_{i=0}^{n-1} \frac{h}{2}\left[f(x_i) + f(x_{i+1})\right] \\ &= \frac{h}{2}\left([f(x_0) + f(x_1)] + [f(x_1) + f(x_2)] + \cdots \right.\\ &\qquad \left. + [f(x_{n-2}) + f(x_{n-1})] + [f(x_{n-1}) + f(x_n)]\right)\end{aligned}$$

On remarque que tous les termes $f(x_i)$ sont répétés deux fois, sauf le premier et le dernier. On en conclut que:

$$\int_a^b f(x)dx \simeq \frac{h}{2}\left(f(x_0) + 2\left[f(x_1) + f(x_2) + \cdots + f(x_{n-1})\right] + f(x_n)\right) \quad (6.22)$$

qui est la *formule des trapèzes composée*. Qu'en est-il du terme d'erreur? Dans chacun des n sous-intervalles $[x_i, x_{i+1}]$, on commet une erreur liée à la méthode du trapèze. Puisque:

$$h = \frac{(b-a)}{n} \text{ et donc } n = \frac{(b-a)}{h}$$

l'erreur totale commise est:

$$n\left(-\frac{f''(\eta)}{12}h^3\right) = -\frac{(b-a)}{h}\frac{f''(\eta)}{12}h^3 = -\frac{(b-a)}{12}f''(\eta)h^2$$

Remarque 6.8

Le raisonnement précédent n'est pas parfaitement rigoureux, même si le résultat final est juste. En effet, dans chaque sous-intervalle $[x_i, x_{i+1}]$, l'erreur liée à la méthode du trapèze simple devrait faire intervenir $f''(\eta_i)$, c'est-à-dire une valeur de η différente pour chaque sous-intervalle. Un autre théorème de la moyenne est alors nécessaire pour conclure (voir Burden et Faires, réf. [2]). L'erreur globale étant donnée par:

$$-\frac{(b-a)}{12}f''(\eta)h^2 \text{ pour } \eta \in [a,b] \quad (6.23)$$

la méthode des trapèzes composée est d'ordre 2. □

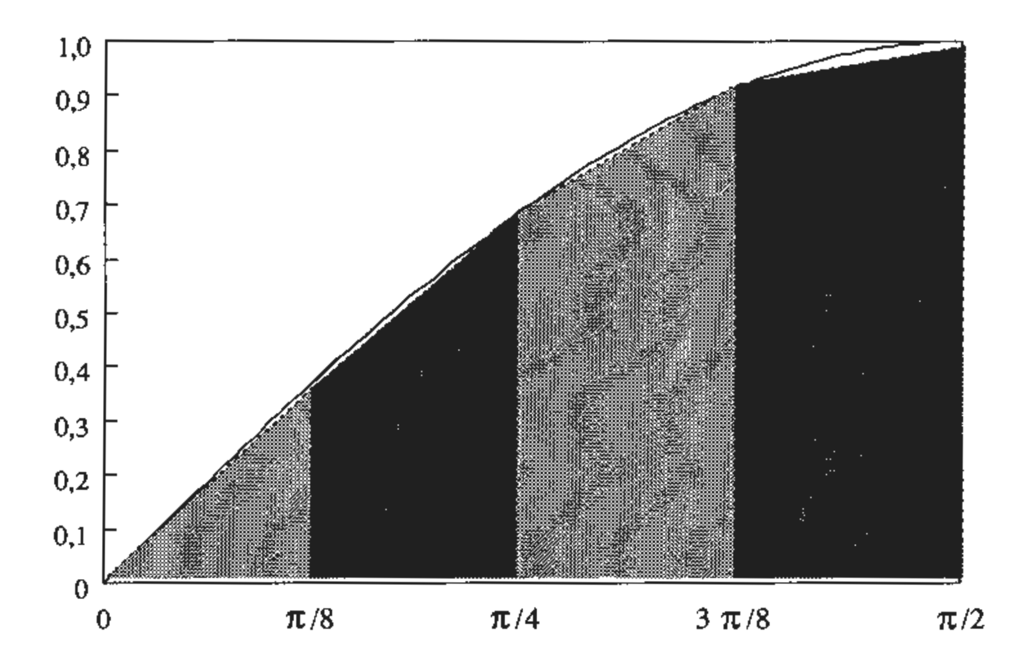


Figure 6.5: Méthode des trapèzes composée, $f(x) = \sin x$ (4 intervalles)

Exemple 6.7
On reprend le calcul de:

$$\int_0^{\frac{\pi}{2}} \sin x dx$$

mais cette fois à l'aide de la méthode des trapèzes composée. Soit d'abord 4 intervalles de longueur:

$$h = \frac{(\frac{\pi}{2} - 0)}{4} = \frac{\pi}{8}$$

tels que les montre la figure 6.5. On a alors:

$$\begin{aligned} \int_0^{\frac{\pi}{2}} \sin x dx &\simeq \frac{\frac{\pi}{8}}{2}(\sin 0 + 2[\sin \frac{\pi}{8} + \sin \frac{\pi}{4} + \sin \frac{3\pi}{8}] + \sin \frac{\pi}{2}) \\ &= 0{,}987\,1158 \end{aligned}$$

soit une erreur absolue d'environ 0,012 88 par rapport à la solution exacte. On constate une nette amélioration en comparaison du résultat obtenu avec

un seul intervalle. Il est intéressant de refaire ce calcul avec 8 intervalles (voir la figure 6.6). La valeur de h est maintenant $\frac{\pi}{16}$ et on a:

$$\begin{aligned}\int_0^{\frac{\pi}{2}} \sin x dx &\simeq \frac{\frac{\pi}{16}}{2}(\sin 0 + 2[\sin \frac{\pi}{16} + \sin \frac{\pi}{8} + \sin \frac{3\pi}{16} \\ &+ \sin \frac{\pi}{4} + \sin \frac{5\pi}{16} + \sin \frac{3\pi}{8} + \sin \frac{7\pi}{16}] + \sin \frac{\pi}{2}) \\ &= 0{,}996\,7852\end{aligned}$$

L'erreur absolue a été réduite à 0,0032. Cette erreur absolue est environ 4 fois plus petite que l'erreur obtenue avec 4 intervalles, ce qui confirme que cette méthode est d'ordre 2. On peut de plus utiliser l'extrapolation de Richardson pour améliorer la précision de ces deux résultats. En utilisant l'équation 6.16 avec $n = 2$, on obtient l'approximation d'ordre au moins 3 suivante:

$$\int_0^{\frac{\pi}{2}} \sin x dx \simeq \frac{2^2(0{,}996\,7852) - 0{,}987\,1158}{2^2 - 1} = 1{,}000\,008\,33$$

ce qui s'approche de plus en plus de la valeur exacte. Comme cela est démontré un peu plus loin, il s'agit en fait d'une approximation d'ordre 4.

● ● ● ● ●

Remarque 6.9
La méthode du trapèze avec un seul intervalle est également connue sous le nom de méthode des trapèzes simple. □

Remarque 6.10
La méthode des trapèzes composée est d'ordre 2. La méthode des trapèzes simple, bien que d'ordre 3, est rarement utilisée car elle est trop imprécise. □

Remarque 6.11
La méthode des trapèzes composée donne un résultat exact si la fonction $f(x)$ est un polynôme de degré inférieur ou égal à 1. Cela s'explique par la présence de la dérivée seconde de $f(x)$ dans le terme d'erreur: celle-ci s'annule dans le cas de polynômes de degré 1. □

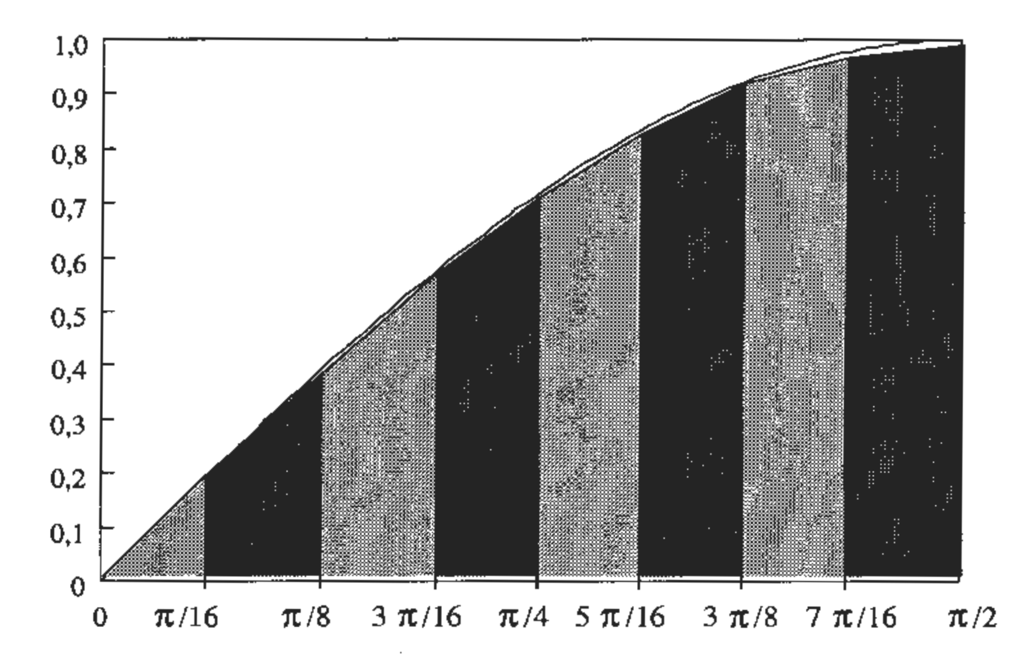


Figure 6.6: Méthode des trapèzes composée, $f(x) = \sin x$ (8 intervalles)

Définition 6.2
Les formules d'intégration numérique sont également appelées *formules de quadrature*.

Définition 6.3
Le *degré de précision* d'une formule de quadrature est la valeur maximale de n pour laquelle cette formule de quadrature intègre exactement tout polynôme de degré inférieur ou égal à n.

Remarque 6.12
Le degré de précision de la formule des trapèzes est 1. □

Formule de Simpson 1/3

Reprenons le raisonnement utilisé avec la méthode des trapèzes, mais cette fois en utilisant un polynôme de degré 2 dont la courbe passe par les points $(x_0, f(x_0))$, $(x_1, f(x_1))$ et $(x_2, f(x_2))$. Ce polynôme est donné par la formule de Newton:

$$p_2(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)$$

On se sert ensuite de l'approximation:

$$\int_{x_0}^{x_2} f(x)dx \simeq \int_{x_0}^{x_2} p_2(x)dx$$

$$= \int_{x_0}^{x_2} \{f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)\}dx$$

On se place de nouveau dans le cas où les abscisses sont également distancées. On pose encore $(x - x_0)/h = s$, ce qui entraîne que $(x - x_i) = (s - i)h$. La dernière expression devient:

$$\int_0^2 \{(f(x_0) + f[x_0, x_1]hs + f[x_0, x_1, x_2]h^2 s(s - 1))\}hds$$

$$= \frac{h}{3}(f(x_0) + 4f(x_1) + f(x_2))$$

où on a remplacé les différences divisées par leur valeur respective:

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{h} \quad \text{et} \quad f[x_0, x_1, x_2] = \frac{f(x_2) - 2f(x_1) + f(x_0)}{2h^2}$$

En résumé, on a:

$$\int_{x_0}^{x_2} f(x)dx \simeq \frac{h}{3}(f(x_0) + 4f(x_1) + f(x_2))$$

qui est la *formule de Simpson 1/3 simple*. Cette terminologie est due au facteur de 1/3 qui multiplie h.

L'analyse de l'erreur est plus délicate dans ce cas. Les numériciens se sont vite rendu compte que cette méthode était plus précise qu'ils ne l'escomptaient. Une analyse plus fine est donc nécessaire. Cette méthode est basée

sur l'utilisation d'un polynôme de degré 2 et on devrait s'attendre à ce que l'erreur soit donnée par:

$$\int_{x_0}^{x_2} E_2(x)dx$$

On peut pousser plus loin l'analyse de l'erreur en introduisant un quatrième point $(x_3, f(x_3))$ quelconque et le polynôme de degré 3 correspondant:

$$p_3(x) = p_2(x) + \frac{(f(x_3) - p_2(x_3))}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)}(x - x_0)(x - x_1)(x - x_2)$$

qui n'est rien d'autre que le polynôme de degré 2 déjà utilisé auquel on ajoute une correction de degré 3 permettant au polynôme de passer également par le point $(x_3, f(x_3))$. Or:

$$\int_{x_0}^{x_2} (x - x_0)(x - x_1)(x - x_2)dx = \int_0^2 s(s-1)(s-2)h^4 ds = 0$$

comme on peut le vérifier facilement. Il s'ensuit que:

$$\int_{x_0}^{x_2} p_2(x)dx = \int_{x_0}^{x_2} p_3(x)dx$$

En utilisant un polynôme de degré 2, on obtient en fait la même précision qu'avec un polynôme de degré 3. Le terme d'erreur est donc de ce fait:

$$\int_{x_0}^{x_2} E_3(x)dx = \int_{x_0}^{x_2} \frac{f''''(\xi)}{4!}(x - x_0)(x - x_1)(x - x_2)(x - x_3)dx$$

Il n'est pas possible à ce stade-ci d'appliquer le théorème de la moyenne, comme nous l'avons fait pour la méthode du trapèze. En effet, la fonction $(x-x_0)(x-x_1)(x-x_2)(x-x_3)$ peut changer de signe dans l'intervalle $[x_0, x_2]$, à moins de choisir judicieusement x_3. Comme le choix de x_3 est arbitraire, on peut poser $x_3 = x_1$. Le terme d'erreur devient alors:

$$\begin{aligned}\int_{x_0}^{x_2} E_3(x)dx &= \int_{x_0}^{x_2} \frac{f''''(\xi)}{4!}(x - x_0)(x - x_1)(x - x_2)(x - x_1)dx \\ &= \int_0^2 \frac{f''''(\xi)}{4!} s(s-1)^2(s-2)h^5 ds\end{aligned}$$

On remarque que la fonction $s(s-1)^2(s-2)$ ne change pas de signe dans l'intervalle $[0, 2]$. La figure 6.7 illustre cette fonction.

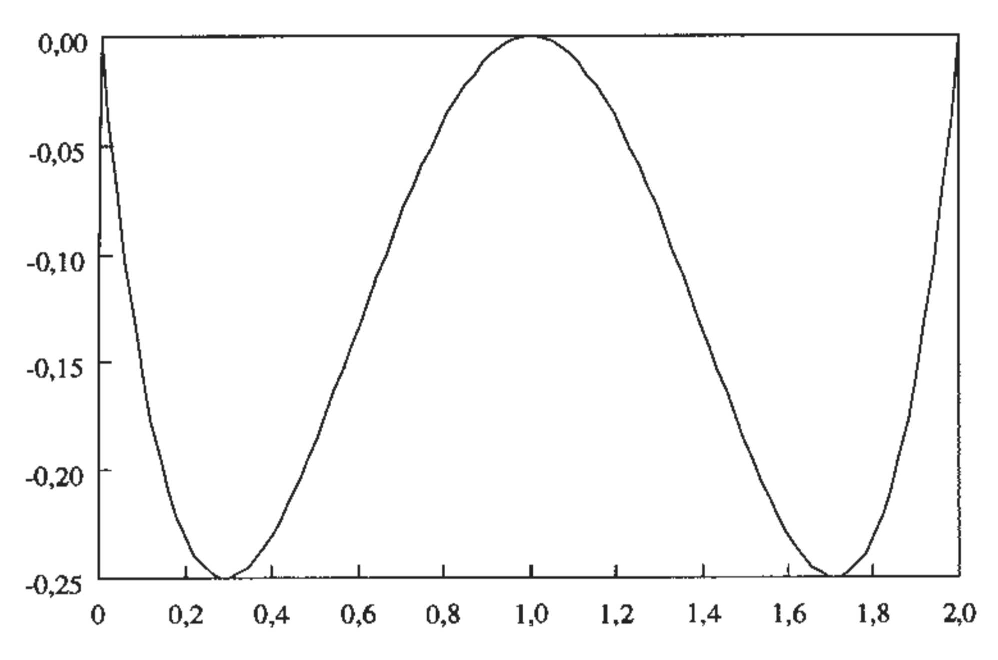


Figure 6.7: Fonction $s(s-1)^2(s-2)$

On peut maintenant se servir du théorème de la moyenne pour obtenir:

$$\int_{x_0}^{x_2} E_3(x)dx = \frac{f''''(\eta)}{4!}h^5 \int_0^2 s(s-1)^2(s-2)ds = -\frac{f''''(\eta)}{90}h^5$$

La méthode de Simpson 1/3 simple se résume donc à:

$$\int_{x_0}^{x_2} f(x)dx = \frac{h}{3}(f(x_0) + 4f(x_1) + f(x_2)) - \frac{f''''(\eta)}{90}h^5 \qquad (6.24)$$

où $\eta \in [x_0, x_2]$.

Remarque 6.13
La valeur de h exprime toujours la distance entre les points x_i, c'est-à-dire qu'elle équivaut dans ce cas à la longueur de l'intervalle divisée par 2. □

La méthode de Simpson 1/3 simple est peu précise, tout comme la méthode du trapèze, comme en témoigne l'exemple suivant.

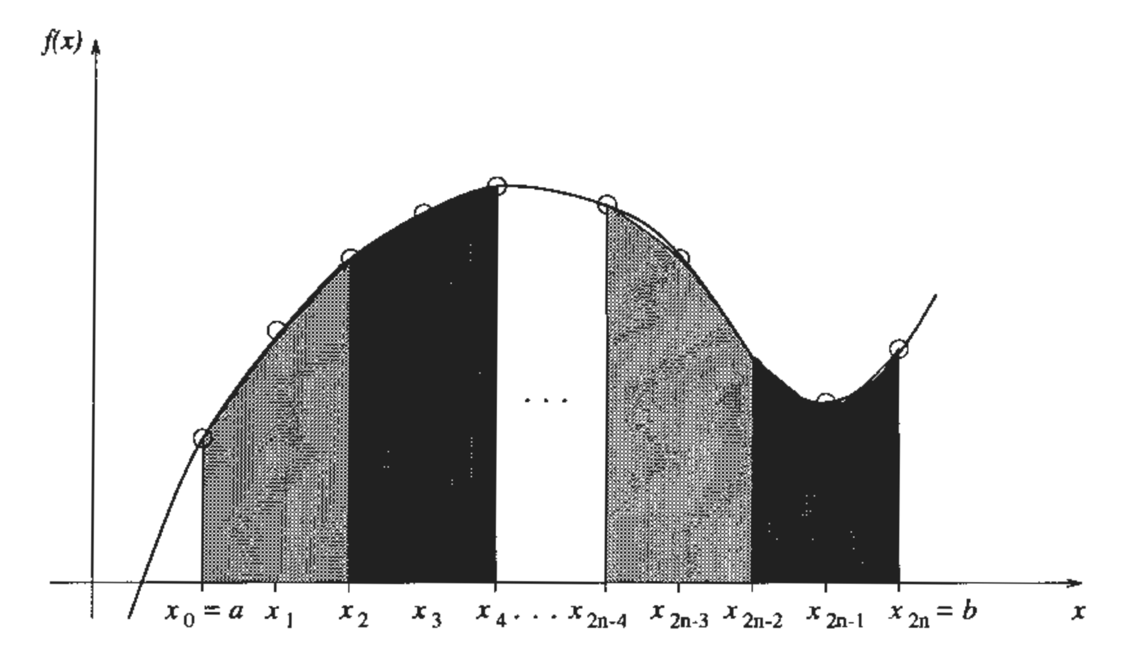


Figure 6.8: Méthode de Simpson 1/3 composée

Exemple 6.8
On reprend une fois de plus le calcul des exemples précédents. Pour la fonction $f(x) = \sin x$ dans l'intervalle $[0, \frac{\pi}{2}]$, on a:

$$\int_0^{\frac{\pi}{2}} \sin x dx \simeq \frac{\frac{\pi}{4}}{3}(\sin 0 + 4\sin\frac{\pi}{4} + \sin\frac{\pi}{2}) = 1{,}002\,2799$$

Ce résultat est plus précis que l'approximation obtenue par la méthode du trapèze simple, mais il demeure peu satisfaisant.

●●●●●

On peut encore une fois améliorer la précision de la formule de Simpson 1/3 en la composant. Puisque la méthode simple requiert deux intervalles, il semble souhaitable de diviser l'intervalle d'intégration $[a, b]$ en $2n$ sous-intervalles et d'utiliser la méthode de Simpson 1/3 simple dans chaque paire de sous-intervalles. La figure 6.8 illustre cette approche. On a alors:

$$\begin{aligned}
\int_a^b f(x)dx &= \sum_{i=0}^{n-1} \int_{x_{2i}}^{x_{2i+2}} f(x)dx \simeq \sum_{i=0}^{n-1} \frac{h}{3}(f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})) \\
&= \frac{h}{3}((f(x_0) + 4f(x_1) + f(x_2)) + (f(x_2) + 4f(x_3) + f(x_4)) + \cdots \\
&\quad + (f(x_{2n-4}) + 4f(x_{2n-3}) + f(x_{2n-2})) \\
&\quad + (f(x_{2n-2}) + 4f(x_{2n-1}) + f(x_{2n}))) \\
&= \frac{h}{3}(f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \cdots \\
&\quad + 4f(x_{2n-3}) + 2(f(x_{2n-2}) + 4f(x_{2n-1}) + f(x_{2n}))
\end{aligned}$$

Tous les termes de rang impair sont mnltipliés par 4 tandis que ceux de rang pair sont multipliés par 2, sauf le premier ($f(x_0)$) et le dernier ($f(x_{2n})$).

L'analyse de l'erreur liée à la méthode de Simpson 1/3 composée est similaire à celle qui s'applique à la méthode des trapèzes composée. En divisant $[a, b]$ en $2n$ intervalles, on utilise n fois la méthode de Simpson 1/3 simple et on commet donc n fois l'erreur liée à cette méthode. On a alors:

$$h = \frac{b-a}{2n} \text{ et donc } n = \frac{b-a}{2h}$$

et l'erreur totale est:

$$n\left(-\frac{f''''(\eta)}{90}h^5\right) = \frac{(b-a)}{2h}(-\frac{f''''(\eta)}{90}h^5) = -\frac{(b-a)}{180}f''''(\eta)h^4$$

Remarque 6.14
Le terme d'erreur de la méthode de Simpson 1/3 composée est:

$$-\frac{(b-a)}{180}f''''(\eta)h^4 \text{ pour nn certain } \eta \in [a, b] \tag{6.25}$$

ce qui en fait une *méthode d'ordre* 4. De plus, en raison de la présence de la dérivée quatrième de $f(x)$, cette méthode est exacte dans le cas des polynômes de degré 3. *Le degré de précision de cette méthode est donc 3.* □

Exemple 6.9
On divise l'intervalle $[0, \frac{\pi}{2}]$ en 4 sous-intervalles de longueur $h = \frac{\pi}{8}$. On a alors:

$$\begin{aligned}\int_0^{\frac{\pi}{2}} \sin x dx &\simeq \frac{\frac{\pi}{8}}{3}(\sin 0 + 4\sin\frac{\pi}{8} + 2\sin\frac{\pi}{4} + 4\sin\frac{3\pi}{8} \\ &+ \sin\frac{\pi}{2}) = 1{,}000\,1346\end{aligned}$$

Pour une quantité de travail similaire, on obtient une précision supérieure à celle de la méthode des trapèzes. Avec 8 sous-intervalles de longueur $\frac{\pi}{16}$, on a:

$$\begin{aligned}\int_0^{\frac{\pi}{2}} \sin x dx &\simeq \frac{\frac{\pi}{16}}{3}(\sin 0 + 4\sin\frac{\pi}{16} + 2\sin\frac{\pi}{8} + 4\sin\frac{3\pi}{16} \\ &+ 2\sin\frac{\pi}{4} + 4\sin\frac{5\pi}{16} + 2\sin\frac{3\pi}{8} + 4\sin\frac{7\pi}{16} \\ &+ \sin\frac{\pi}{2}) = 1{,}000\,008\,296\end{aligned}$$

Cette plus grande précision vient du fait que cette méthode est d'ordre 4. On constate qu'en passant de 4 à 8 intervalles (c'est-à-dire en divisant h par 2) on divise l'erreur par un facteur d'environ 16,22, ce qui confirme l'ordre 4 de la méthode. On peut également utiliser l'extrapolation de Richardson 6.16 avec $n = 4$ à partir de ces deux valeurs. On obtient ainsi l'approximation:

$$\frac{2^4(1{,}000\,008\,296) - 1{,}000\,1346}{2^4 - 1} = 0{,}999\,999\,876$$

qui est d'ordre au moins 5. On verra plus loin qu'elle est en fait d'ordre 6.

• • • • •

Exemple 6.10
On doit calculer:

$$\int_0^1 e^{-x^2} dx$$

à l'aide de la méthode de Simpson 1/3 composée avec 8 intervalles de longueur:

$$\frac{1-0}{8} = \frac{1}{8}$$

Comme la fonction e^{-x^2} n'a pas de primitive, il faut absolument utiliser une méthode numérique. Dans ce cas:

$$\begin{aligned}\int_0^1 e^{-x^2}dx &\simeq \frac{\frac{1}{8}}{3}(e^0 + 4e^{-0,125^2} + 2e^{-0,25^2} + 4e^{-0,375^2} + 2e^{-0,5^2} \\ &\quad +4e^{-0,625^2} + 2e^{-0,75^2} + 4e^{-0,875^2} + e^{-1,0}) \\ &= 0,746\,826\,1205\end{aligned}$$

Il est intéressant de poursuivre les calculs un peu plus loin et de comparer une fois de plus les méthodes des trapèzes et de Simpson 1/3 composées. En prenant 64 intervalles et en travaillant en double précision (IEEE), on obtient les valeurs suivantes.

Méthode des trapèzes composée	0,746 809 163
Méthode de Simpson 1/3 composée	0,746 824 133
Solution exacte à 9 chiffres	0,746 824 133

ce qui démontre la supériorité de la méthode de Simpson.

• • • • •

On peut poursuivre dans la même voie et développer des formules de Newton-Cotes basées sur des polynômes de degré de plus en plus élevé. Nous ne présentons ci-dessous que les formules de Simpson 3/8 et de Boole sans les démontrer.

Formule de Simpson 3/8

Si on utilise un polynôme de degré 3 dans l'intervalle $[x_0, x_3]$ et passant par les points $((x_i, f(x_i))$ pour $i = 0, 1, 2, 3)$, on obtient la formule de Simpson 3/8 simple qui s'écrit:

$$\int_{x_0}^{x_3} f(x)dx = \frac{3h}{8}(f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)) - \frac{3f''''(\eta)}{80}h^5 \quad (6.26)$$

pour un certain $\eta \in [x_0, x_3]$.

On peut également composer cette méthode en divisant l'intervalle d'intégration $[a, b]$ en $3n$ sous-intervalles de longueur:

$$h = \frac{b-a}{3n}$$

et en utilisant la formule de Simpson 3/8 simple dans chaque triplet de sous-intervalles. On obtient alors:

$$\begin{aligned}
\int_a^b f(x)dx &= \sum_{i=0}^{n-1} \int_{x_{3i}}^{x_{3i+3}} f(x)dx \\
&\simeq \sum_{i=0}^{n-1} \frac{3h}{8}(f(x_{3i}) + 3f(x_{3i+1}) + 3f(x_{3i+2}) + f(x_{3i+3})) \\
&= \frac{3h}{8}(f(x_0) + 3f(x_1) + 3f(x_2) + 2f(x_3) + 3f(x_4) + \cdots \\
&\quad + 2f(x_{3n-3}) + 3f(x_{3n-2}) + 3f(x_{3n-1}) + f(x_{3n}))
\end{aligned}$$

et le terme d'erreur:

$$n\left(-\frac{3f''''(\eta)}{80}h^5\right) = -\frac{(b-a)}{3h}\frac{3f''''(\eta)}{80}h^5 = -\frac{(b-a)f''''(\eta)}{80}h^4$$

Remarque 6.15
La méthode de Simpson 3/8 composée a le même ordre de convergence (4) et le même degré de précision (3) que la méthode de Simpson 1/3 composée. Pour cette raison, on lui préfère sonvent la méthode de Simpson 1/3. □

Formule de Boole

Si on a au départ un polynôme de degré 4 dans l'intervalle $[x_0, x_4]$ dont la courbe passe par les points $((x_i, f(x_i))$ pour $i = 0, 1, 2, 3, 4)$, la formule de Boole simple s'écrit:

$$\begin{aligned}
\int_{x_0}^{x_4} f(x)dx &= \frac{2h}{45}[7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)] \\
&\quad - \frac{8f^{(6)}(\eta)}{945}h^7
\end{aligned} \tag{6.27}$$

pour un certain $\eta \in [x_0, x_4]$.

On compose cette méthode en divisant cette fois l'intervalle d'intégration $[a, b]$ en $4n$ sous-intervalles de longueur:

$$h = \frac{b-a}{4n}$$

et en utilisant la formule de Boole simple dans chaque quadruplet de sous-intervalles. On obtient alors:

$$\begin{aligned}
\int_a^b f(x)dx &= \sum_{i=0}^{n-1} \int_{x_{4i}}^{x_{4i+4}} f(x)dx \\
&\simeq \sum_{i=0}^{n-1} \frac{2h}{45}(7f(x_{4i}) + 32f(x_{4i+1}) + 12f(x_{4i+2}) \\
&\quad + 32f(x_{4i+3}) + 7f(x_{4i+4})) \\
&= \frac{2h}{45}(7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 14f(x_4) + \cdots \\
&\quad + 32f(x_{4n-5}) + 14f(x_{4n-4}) + 32f(x_{4n-3}) + 12f(x_{4n-2}) \\
&\quad + 32f(x_{4n-1}) + 7f(x_{4n}))
\end{aligned}$$

et le terme d'erreur:

$$n\left(-\frac{8f^{(6)}(\eta)}{945}h^7\right) = -\frac{(b-a)}{4h}\frac{8f^{(6)}(\eta)}{945}h^7 = -\frac{2(b-a)f^{(6)}(\eta)}{945}h^6$$

Remarque 6.16
En ce qui concerne l'erreur, il se produit un phénomène déjà observé avec la formule de Simpson 1/3 en ce sens qne la formule de Boole conduit à une approximation d'ordre 6 au lieu de 5. La méthode de Boole a de plus un degré de précision de 5 puisqu'elle est exacte pour tous les polynômes de degré inférieur ou égal à 5. □

6.4.2 Méthode de Romberg

La méthode de Romberg est une méthode d'intégration qui permet d'atteindre des résultats très précis. Elle est basée sur une utilisation très astucieuse de la méthode des trapèzes composée (d'ordre 2) et de la technique d'extrapolation de Richardson 6.16. On peut en effet démontrer, sous des hypothèses de régularité suffisante de la fonction $f(x)$, que le terme d'erreur de la méthode des trapèzes composée s'écrit:

$$-\frac{(b-a)}{12}f''(\eta)h^2 = c_2h^2 + c_4h^4 + c_6h^6 + \cdots$$

où les c_i sont des constantes. L'information supplémentaire que l'on tire de cette relation est que seuls les termes d'ordre pair sont présents. L'absence des puissances impaires de h permet, du point de vue de l'extrapolation de Richardson, de gagner deux ordres de convergence à chaque extrapolation. De plus, les valeurs extrapolées, qui sont d'ordre 4, peuvent à leur tour être extrapolées pour passer à l'ordre 6, et ainsi de suite. Cette utilisation systématique de l'extrapolation de Richardson permet d'obtenir successivement des approximations de:

$$\int_a^b f(x)dx$$

d'ordre $2, 4, 6, 8$ et plus. Sur le plan pratique, on obtient généralement des résultats extrêmement précis.

Dans un premier temps, introduisons quelques notations. *On note $T_{1,i}$ le résultat obtenu à l'aide de la méthode des trapèzes composée avec 2^{i-1} intervalles.* Les $T_{1,i}$ sont des approximations d'ordre 2. Pour passer de $T_{1,i}$ à $T_{1,i+1}$, on doit doubler le nombre de sous-intervalles, ce qui revient à diviser la valeur de h par deux. Au moyen de l'extrapolation de Richardson 6.16 avec $n = 2$, on définit alors:

$$T_{2,i} = \frac{2^2T_{1,i+1} - T_{1,i}}{2^2 - 1} \tag{6.28}$$

et les $T_{2,i}$ sont des approximations d'ordre 4. On pose ensuite successivement:

$$
\begin{aligned}
T_{3,i} &= \frac{2^4 T_{2,i+1} - T_{2,i}}{2^4 - 1} \\
T_{4,i} &= \frac{2^6 T_{3,i+1} - T_{3,i}}{2^6 - 1} \\
T_{5,i} &= \frac{2^8 T_{4,i+1} - T_{4,i}}{2^8 - 1} \\
&\vdots
\end{aligned}
\tag{6.29}
$$

ce qui définit un triangle de la forme:

$T_{1,1}$	$T_{1,2}$	$T_{1,3}$	$T_{1,4}$	$T_{1,5}$	$T_{1,6}$	(ordre 2)
$T_{2,1}$	$T_{2,2}$	$T_{2,3}$	$T_{2,4}$	$T_{2,5}$		(ordre 4)
$T_{3,1}$	$T_{3,2}$	$T_{3,3}$	$T_{3,4}$			(ordre 6)
$T_{4,1}$	$T_{4,2}$	$T_{4,3}$				(ordre 8)
$T_{5,1}$	$T_{5,2}$					(ordre 10)
$T_{6,1}$						(ordre 12)

Chaque ligne de ce triangle est de deux ordres de convergence plus précis que la ligne précédente. La première ligne est tout simplement constituée des approximations obtenues à l'aide de la méthode des trapèzes composée avec $1, 2, 4, 8, 16 \cdots$ intervalles. Pour passer d'une ligne à l'autre, on utilise l'extrapolation de Richardson par le biais des relations 6.28 et 6.29.

Remarque 6.17
On peut montrer (voir les exercices de fin de chapitre) que la deuxième ligne de ce tableau n'est autre que le résultat de la méthode de Simpson 1/3 avec respectivement $2, 4, 8 \cdots$ intervalles. On pourrait donc éliminer la première ligne et commencer directement avec la méthode de Simpson. □

Exemple 6.11
On a déjà obtenu, lors de calculs précédents, les valeurs $T_{1,1} = 0{,}785\,3982$, $T_{1,3} = 0{,}987\,1158$ et $T_{1,4} = 0{,}996\,7852$ correspondant à la formule des trapèzes composée avec respectivement 1, 4 et 8 intervalles pour évaluer:

$$\int_0^{\frac{\pi}{2}} \sin x dx$$

Il est alors possible de remplir la première ligne du tableau en calculant:

$$T_{1,2} = \frac{\frac{\pi}{4}}{2}(\sin 0 + 2\sin\frac{\pi}{4} + \sin\frac{\pi}{2}) = 0{,}948\,0594$$

On peut ensuite effectuer les différentes extrapolations de Richardson.

0,785 3982	0,948 0594	0,987 1158	0,996 7852	(ordre 2)
1,002 2799	1,000 1346	1,000 0083		(ordre 4)
0,999 9916	0,999 9999			(ordre 6)
1,000 0000				(ordre 8)

La première ligne du tableau étant d'ordre 2, la deuxième ligne est donnée par:

$$\frac{(2^2)(0{,}948\,0594) - 0{,}785\,3982}{2^2 - 1} = 1{,}002\,2799$$

$$\frac{(2^2)(0{,}987\,1158) - 0{,}948\,0594}{2^2 - 1} = 1{,}000\,1346$$

$$\frac{(2^2)(0{,}996\,7852) - 0{,}987\,1158}{2^2 - 1} = 1{,}000\,0083$$

qui sont toutes des approximations d'ordre 4. La troisième ligne devient alors:

$$\frac{(2^4)(1{,}000\,1346) - 1{,}002\,2799}{2^4 - 1} = 0{,}999\,9916$$

$$\frac{(2^4)(1{,}000\,0083) - 1{,}000\,1346}{2^4 - 1} = 0{,}999\,9999$$

d'ordre 6. Puis enfin:

$$\frac{(2^6)(0{,}999\,9999) - 0{,}999\,9916}{2^6 - 1} = 1{,}000\,0000$$

Il en résulte une approximation d'ordre 8 ayant plus de 7 chiffres significatifs. On remarque que la précision augmente à mesure que l'on se déplace vers le bas (car l'ordre de l'approximation augmente) et vers la droite sur une même ligne (car h est divisé par 2 entre chaque valeur). □

● ● ● ● ●

Exemple 6.12
Soit une fonction $f(x)$ connue seulement pour quelques valeurs de x.

x	$f(x)$
0,00	0,3989
0,25	0,3867
0,50	0,3521
0,75	0,3011
1,00	0,2420

On tente d'évaluer:

$$\int_0^1 f(x)dx$$

selon la méthode de Romberg. Puisqu'il y a en tout 5 points, on peut utiliser la méthode des trapèzes composée avec 1, 2 et 4 intervalles seulement. On a respectivement:

$$T_{1,1} = \frac{1}{2}(0{,}3989 + 0{,}2420) = 0{,}320\,45$$

$$T_{1,2} = \frac{\frac{1}{2}}{2}(0{,}3989 + 2(0{,}3521) + 0{,}2420) = 0{,}336\,275$$

$$T_{1,3} = \frac{\frac{1}{4}}{2}(0{,}3989 + 2(0{,}3867 + 0{,}3521 + 0{,}3011) + 0{,}2420)$$

$$= 0{,}340\,0875$$

On peut dès lors remplir la première ligne du tableau de la méthode de Romberg.

0,320 45	0,336 275	0,340 0875	(ordre 2)
0,341 55	0,341 3583		(ordre 4)
0,341 3456			(ordre 6)

Les autres lignes du tableau sont tirées elles aussi des relations 6.28 et 6.29:

$$\frac{(2^2)(0{,}336\,275) - 0{,}320\,45}{2^2 - 1} = 0{,}341\,55$$

$$\frac{(2^2)(0{,}340\,0875) - 0{,}336\,275}{2^2 - 1} = 0{,}341\,3583$$

$$\frac{(2^4)(0{,}341\,3583) - 0{,}341\,55}{2^4 - 1} = 0{,}341\,3456$$

On obtient ainsi une approximation d'ordre 6 de l'intégrale.

● ● ● ● ●

Remarque 6.18
Dans le cas d'une fonction connue seulement en certains points, comme dans l'exemple précédent, le nombre de points doit être de la forme 2^n+1 pour que la méthode de Romberg puisse s'appliquer. En effet, il faut que le nombre de sous-intervalles soit une puissance de 2. Dans l'exemple précédent, on avait $2^2 + 1$ points et 4 intervalles. □

6.4.3 Quadratures de Gauss

Les quadratures de Gauss reposent sur un raisonnement différent de celui qui est à la base des méthodes de Newton-Cotes. D'une certaine façon, on cherche à optimiser les schémas d'intégration numérique en choisissant plus judicieusement les points où est évaluée la fonction $f(x)$. Dans le cas où l'évaluation de $f(x)$ est coûteuse en temps de calcul, ces quadratures permettent d'atteindre une grande précision avec relativement peu d'évaluations de $f(x)$. Par exemple, la méthode du trapèze requiert l'évaluation de

la fonction $f(x)$ aux deux extrémités de l'intervalle sous la forme:

$$\int_a^b f(x)dx \simeq \frac{(b-a)}{2}(f(a)+f(b))$$

Nous avons vu que le degré de précision de cette méthode est 1, car cette quadrature est exacte dans le cas de tout polynôme de degré inférieur ou égal à 1. On peut se demander s'il est possible de trouver deux points situés dans l'intervalle d'intégration ainsi que des coefficients appropriés de telle sorte que l'expression:

$$\int_a^b f(x)dx \simeq w_1 f(t_1) + w_2 f(t_2)$$

ait un degré de précision supérieur à celui de la méthode du trapèze. Bien sûr, si:

$$w_1 = w_2 = \frac{(b-a)}{2}, \quad t_1 = a \text{ et } t_2 = b$$

on retrouve la formule du trapèze. Mais est-ce un choix optimal?

Pour répondre à cette question, nous allons dans un premier temps nous restreindre à l'intervalle $[-1, 1]$, où nous ferons tout le développement. Pour un intervalle quelconque, il suffira d'effectuer le changement de variable:

$$x = \frac{(b-a)t + (a+b)}{2} \text{ et } dx = \frac{(b-a)}{2}dt \tag{6.30}$$

qui envoie l'intervalle $[-1, 1]$ sur un intervalle quelconque $[a, b]$. En effet, le changement de variable 6.30 permet d'écrire que:

$$\int_a^b f(x)dx = \int_{-1}^1 f\left(\frac{(b-a)t + (a+b)}{2}\right)\frac{(b-a)}{2}dt = \frac{(b-a)}{2}\int_{-1}^1 g(t)dt$$

où:

$$g(t) = f\left(\frac{(b-a)t + (a+b)}{2}\right)$$

Il est donc toujours possible de revenir à l'intervalle $[-1, 1]$. De manière générale, on cherche des expressions de la forme:

$$\int_{-1}^1 g(t)dt \simeq \sum_{i=1}^n w_i g(t_i) \tag{6.31}$$

dont le degré de précision soit le plus élevé possible.

Définition 6.4
L'expression 6.31 est appelée *quadrature de Gauss à n points.* Les t_i sont appelés *points d'intégration*, tandis que les coefficients w_i sont les *poids d'intégration.*

On choisit les points et les poids d'intégration de façon à ce que la quadrature 6.31 soit exacte dans le cas des polynômes de degré le plus élevé possible. Puisque tout polynôme de degré n peut s'écrire:

$$p_n(t) = \sum_{i=0}^{n} c_n t^n$$

il suffit que la relation 6.31 soit exacte successivement pour $g(t) = t^k$, pour $k = 0, 1, 2, \cdots, n$. On gagne à accroître le plus possible l'exposant k. Le degré maximal atteint dépend du nombre de points n. *Puisqu'il y a* $2n$ *coefficients à déterminer dans l'équation 6.31, il est raisonnable de penser que l'on peut atteindre le degré* $(2n - 1)$. La valeur de k varie donc entre 0 et $2n - 1$.

Quadrature de Gauss à 1 point

Cherchons donc une expression de la forme:

$$\int_{-1}^{1} g(t)dt = w_1 g(t_1) \tag{6.32}$$

qui soit exacte dans le cas des polynômes de degré le plus élevé possible. Commençons par les polynômes de degré 0. La formule 6.32 doit être exacte pour $g(t) = 1$, ce qui donne une première équation:

$$\int_{-1}^{1} 1dt = 2 = w_1$$

et l'unique poids d'intégration est déjà déterminé. L'équation 6.31 doit de plns être exacte pour $g(t) = t$. On trouve donc:

$$\int_{-1}^{1} tdt = 0 = w_1 t_1 = 2t_1$$

ce qui entraîne que $t_1 = 0$. Ainsi, la *quadrature de Gauss à 1 point* s'écrit:

$$\int_{-1}^{1} g(t)dt \simeq 2g(0)$$

et est exacte pour tout polynôme de degré 1.

Remarque 6.19
La quadrature de Gauss à 1 point a le même degré de précision (1) que la méthode du trapèze, qui est une formule à 2 poiuts. La quadrature de Gauss à 1 point est également connue sous le nom de *formule du point milieu.* □

Quadrature de Gauss à 2 points

On doit maintenant déterminer les 4 coefficieuts inconnus de l'expression:

$$\int_{-1}^{1} g(t)dt \simeq w_1 g(t_1) + w_2 g(t_2) \tag{6.33}$$

On remarque immédiatement que t_1 doit être différent de t_2 et que les deux w_i doivent être non nuls. Sinon, on se retrouve avec une formule à 1 point. Il nous faut alors 4 équations qui proviendront de la relation 6.33, où on choisit successivement $g(t) = 1$, $g(t) = t$, $g(t) = t^2$ et $g(t) = t^3$. Les 4 équations résultantes sont:

$$\int_{-1}^{1} 1dt = 2 \quad = \quad w_1 + w_2 \tag{6.34}$$

$$\int_{-1}^{1} tdt = 0 \quad = \quad w_1 t_1 + w_2 t_2 \tag{6.35}$$

$$\int_{-1}^{1} t^2 dt = \frac{2}{3} \quad = \quad w_1 t_1^2 + w_2 t_2^2 \tag{6.36}$$

$$\int_{-1}^{1} t^3 dt = 0 \quad = \quad w_1 t_1^3 + w_2 t_2^3 \tag{6.37}$$

et forment un système non linéaire qu'il est heureusement possible de résoudre analytiquement. On multiplie l'équatiou 6.35 par t_1^2 et on soustrait du résultat l'équation 6.37 pour obtenir:

$$w_2 t_2 (t_1^2 - t_2^2) = 0$$

Pour que ce produit soit nul, il faut que l'un ou l'autre des facteurs s'annule, c'est-à-dire:

- $w_2 = 0$.

  Cette possibilité doit être écartée, car dans ce cas la formule de Gauss à 2 points 6.33 dégénère en une formule à 1 seul point.

- $t_2 = 0$.

  De l'équation 6.35, on tire que $w_1 = 0$ ou $t_1 = 0$, ce qui conduit de nouveau à une formule à 1 point.

- $t_1^2 = t_2^2$.

  On en conclut que $t_1 = -t_2$, puisque le cas $t_1 = t_2$ conduit encore à une formule à 1 point.

Cette conclusion permet d'obtenir les poids d'intégration. En effet, en vertu de l'équation 6.35:

$$t_1(w_1 - w_2) = 0$$

et puisque t_1 ne peut être nul, $w_1 = w_2$ et la relation 6.34 entraîne que:

$$w_1 = w_2 = 1$$

Enfin, selon l'équation 6.36, on a:

$$\frac{2}{3} = t_1^2 + t_2^2 = t_1^2 + (-t_1)^2 = 2t_1^2$$

ce qui entraîne que:

$$t_1 = -\sqrt{\frac{1}{3}} \text{ et donc } t_2 = \sqrt{\frac{1}{3}}$$

La *formule de Gauss à 2 points* s'écrit donc:

$$\int_{-1}^{1} g(t)dt \simeq g\left(-\sqrt{\frac{1}{3}}\right) + g\left(\sqrt{\frac{1}{3}}\right)$$

et est exacte dans le cas des polynômes de degré inférieur ou égal à 3.

Remarque 6.20
Pour un même nombre de points d'intégration, la quadrature de Gauss à 2 points a un degré de précision de 3 par comparaison avec 1 pour la méthode du trapèze. Pour un même effort de calcul, on a ainsi une plus grande précision. □

Quadratures de Gauss à n points

Sans entrer dans les détails, il est possible de déterminer des quadratures de Gauss avec un grand nombre de points. Ces quadratures sont particulièrement efficaces et sont utilisées, par exemple, dans la méthode des éléments finis (voir Reddy, réf. [24]). On détermine les $2n$ coefficients w_i et t_i en résolvant un système non linéaire de $2n$ éqnations que l'on obtient en prenant $g(t) = t^k$ pour $k = 0, 1, 2, \cdots, (2n-1)$.

On peut également démontrer que les points d'intégration de Gauss sont les racines des polynômes de Legendre définis par:

$$L_0(x) = 1 \text{ et } L_1(x) = x$$

et par la formule de récurrence:

$$(n+1)L_{n+1}(x) = (2n+1)xL_n(x) - nL_{n-1}(x)$$

Il est alors facile de démontrer que:

$$L_2(x) = \frac{1}{2}(3x^2 - 1)$$

dont les racines sont $\pm\sqrt{1/3}$. En résumé, on a le résultat général suivant.

Théorème 6.2

La quadrature de Gauss à n points 6.31 est exacte dans le cas des polynômes de degré $(2n-1)$. Le degré de précision de cette quadrature est donc $(2n-1)$.

Le terme d'erreur est donné par:

$$\frac{2^{2n+1}(n!)^4}{(2n+1)((2n)!)^3} f^{(2n)}(\xi) \text{ où } \xi \in [-1, 1] \quad \square \tag{6.38}$$

Le tableau suivant résume les principales quadratures de Gauss (voir Burden et Faires, réf. [2]; Chapra et Canale, réf. [4]; et Gerald et Wheatly, réf. [12]).

n	Points d'intégration t_i	Poids d'intégration w_i	Degré de précision
1	0	2	1
2	−0,577 350 2629	1	3
	+0,577 350 2629	1	
3	−0,774 596 669	0,555 555 556	5
	0,0	0,888 888 889	
	+0,774 596 669	0,555 555 556	
4	−0,861 136 312	0,347 854 845	7
	−0,339 981 044	0,652 145 155	
	+0,339 981 044	0,652 145 155	
	+0,861 136 312	0,347 854 845	
5	−0,906 179 846	0,236 926 885	9
	−0,538 469 310	0,478 628 670	
	0,0	0,568 888 889	
	+0,538 469 310	0,478 628 670	
	+0,906 179 846	0,236 926 885	

Exemple 6.13

On doit évaluer:

$$\int_0^1 (4x^3 + 3x^2 + 2)dx$$

dont la valeur exacte est 4. Il faut d'abord effectuer le changement de variable 6.30 pour obtenir:

$$\int_0^1 (4x^3 + 3x^2 + 2)dx = \frac{1}{2}\int_{-1}^1 \left(4\left(\frac{t+1}{2}\right)^3 + 3\left(\frac{t+1}{2}\right)^2 + 2\right) dt$$

La formule de Gauss à 1 point donne l'approximation:

$$\int_0^1 (4x^3 + 3x^2 + 2)dx \simeq \frac{2}{2}\left(4\left(\frac{0+1}{2}\right)^3 + 3\left(\frac{0+1}{2}\right)^2 + 2\right) = 3{,}25$$

Par contre, la quadrature à 2 points donne:

$$\int_0^1 (4x^3 + 3x^2 + 2)dx$$

$$\begin{aligned}
&\simeq \frac{1}{2}\left[4\left(\frac{-\sqrt{1/3}+1}{2}\right)^3 + 3\left(\frac{-\sqrt{1/3}+1}{2}\right)^2 + 2\right.\\
&\qquad \left. + 4\left(\frac{\sqrt{1/3}+1}{2}\right)^3 + 3\left(\frac{\sqrt{1/3}+1}{2}\right)^2 + 2\right]\\
&= \frac{1}{2}\left[\left(\frac{-\sqrt{1/3}+1}{2}\right)^2\left[4\left(\frac{-\sqrt{1/3}+1}{2}\right) + 3\right] + 2\right.\\
&\qquad \left. + \left(\frac{\sqrt{1/3}+1}{2}\right)^2\left[4\left(\frac{\sqrt{1/3}+1}{2}\right) + 3\right] + 2\right]\\
&= \frac{1}{2}\left[\left(\frac{4/3 - 2\sqrt{1/3}}{4}\right)\left(5 - 2\sqrt{1/3}\right)\right.\\
&\qquad \left. + \left(\frac{4/3 - 2\sqrt{1/3}}{4}\right)\left(5 - 2\sqrt{1/3}\right) + 4\right]\\
&= \frac{1}{2}[4 + 4] = 4
\end{aligned}$$

L'exactitude de ce résultat était prévisible, car la fonction intégrée est de degré 3 et la quadrature de Gauss à 2 points est exacte (par construction) pour tout polynôme de degré inférieur ou égal à 3.

● ● ● ● ●

Exemple 6.14

Grâce au changement de variable 6.30, on a:

$$\int_0^{\frac{\pi}{2}} \sin x dx = \frac{\frac{\pi}{2}}{2}\int_{-1}^{1} \sin\left(\frac{\pi(t+1)}{4}\right) dt$$

La quadrature de Gauss à 2 points donne l'approximation:

$$\begin{aligned}
\int_0^{\frac{\pi}{2}} \sin x dx &\simeq \frac{\pi}{4}\left(\sin\left(\frac{\pi(t_1+1)}{4}\right)+\sin\left(\frac{\pi(t_2+1)}{4}\right)\right) \\
&= \frac{\pi}{4}(\sin(0{,}331\,948\,322)+\sin(1{,}238\,848\,005)) \\
&= 0{,}998\,472\,614
\end{aligned}$$

Si on tient compte du fait que l'on a évalué la fonction $\sin x$ en seulement deux points, ce résultat est d'une précision remarquable. Par ailleurs, la formule à trois points donne:

$$\begin{aligned}
\int_0^{\frac{\pi}{2}} \sin x dx &\simeq \frac{\pi}{4}\left(w_1 \sin\left(\frac{\pi(t_1+1)}{4}\right)+w_2 \sin\left(\frac{\pi(t_2+1)}{4}\right)\right. \\
&\quad \left. + w_3 \sin\left(\frac{\pi(t_3+1)}{4}\right)\right) \\
&= \frac{\pi}{4}((0{,}555\,555\,556)\sin(0{,}177\,031\,362) \\
&\quad + (0{,}888\,888\,889)\,\mathrm{siu}(0{,}785\,398\,164) \\
&\quad + (0{,}555\,555\,556)\sin(1{,}774\,596\,669)) \\
&= 1{,}000\,008\,1821
\end{aligned}$$

La formule de Gauss à 3 points est donc plus précise que la méthode de Simpson 1/3 simple, qui nécessite en outre l'évaluation de la fonction $\sin x$ en trois points. Pour obtenir une précision similaire avec la méthode de Simpson 1/3, nous avons dû utiliser 8 intervalles et donc 9 évaluations de la fonction $\sin x$.

● ● ● ● ●

Exemple 6.15
Soit:

$$\int_0^1 \frac{1}{\sqrt{x}} dx$$

dont la valeur exacte est 2. On remarque immédiatement que, parmi les méthodes proposées, seules les quadratures de Gauss peuvent s'appliquer. En effet, toutes les autres méthodes nécessitent l'évaluation de la fonction $x^{-1/2}$ en $x = 0$, qui n'est pas définie à cet endroit. On peut prévoir que le calcul de cette intégrale sera difficile. Dans un premier temps, le changement de variable 6.30 donne:

$$\int_0^1 \frac{1}{\sqrt{x}}dx = \frac{1}{2}\int_{-1}^1 \frac{1}{\sqrt{\frac{t+1}{2}}}dt = \frac{\sqrt{2}}{2}\int_{-1}^1 \frac{1}{\sqrt{t+1}}dt$$

La formule à 2 points donne:

$$\int_0^1 \frac{1}{\sqrt{x}}dx \simeq \frac{\sqrt{2}}{2}\left(\frac{1}{\sqrt{-\sqrt{1/3}+1}} + \frac{1}{\sqrt{\sqrt{1/3}+1}}\right) = 1{,}650\,680\,13$$

La formule à 3 points est légèrement plus précise. En se servant de la table des quadratures de Gauss pour établir la valeur des différents coefficients, on obtient:

$$\int_0^1 \frac{1}{\sqrt{x}}dx \simeq \frac{\sqrt{2}}{2}\left(w_1\frac{1}{\sqrt{t_1+1}} + w_2\frac{1}{\sqrt{t_2+1}} + w_3\frac{1}{\sqrt{t_3+1}}\right) = 1{,}750\,863\,166$$

La précision demeure insatisfaisante, mais il faut admettre qu'il s'agit d'un problème difficile. Les quadratures de Gauss à 4 ou à 5 points amélioreraient encore la qualité des résultats.

● ● ● ● ●

Remarque 6.21

Les quadratures de Gauss permettent d'évaluer des intégrales avec une grande précision. Toutefois, chaque fois que l'on change l'ordre de la quadrature, les points t_i et les poids w_i d'intégration changent eux aussi. Il devient alors à peu près impossible d'utiliser une technique d'extrapolation comme la méthode de Romberg. □

6.4.4 Intégration à l'aide des splines

Si la spline constitue une bonne approximation d'une fonction $f(x)$ connue seulement en quelques points, elle peut également servir pour calculer l'intégrale de cette fonction. On obtient ainsi une expression voisine de celle

qui caractérise la méthode des trapèzes composée, à laquelle s'ajoute une approximation du terme d'erreur de cette méthode. On pose:

$$\int_a^b f(x)dx \simeq \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} p_i(x)dx$$

où $p_i(x)$ est le polynôme de degré 3 de la spline dans l'intervalle $[x_{i-1}, x_i]$. L'expression de ce polynôme est bien sûr (voir l'équation 5.30):

$$\begin{aligned} p_i(x) &= -f''_{i-1}\frac{(x-x_i)^3}{6h_i} + f''_i\frac{(x-x_{i-1})^3}{6h_i} \\ &\quad - \left(\frac{f(x_{i-1})}{h_i} - \frac{h_i f''_{i-1}}{6}\right)(x-x_i) \\ &\quad + \left(\frac{f(x_i)}{h_i} - \frac{h_i f''_i}{6}\right)(x-x_{i-1}) \end{aligned} \tag{6.39}$$

En intégrant ce polynôme, on obtient:

$$\int_a^b f(x)dx \simeq \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} p_i(x)dx$$

$$= \sum_{i=1}^{n} \left\{ -f''_{i-1}\frac{(x-x_i)^4}{24h_i} + f''_i\frac{(x-x_{i-1})^4}{24h_i} \right.$$

$$\left. - \left(\frac{f(x_{i-1})}{h_i} - \frac{h_i f''_{i-1}}{6}\right)\frac{(x-x_i)^2}{2} + \left(\frac{f(x_i)}{h_i} - \frac{h_i f''_i}{6}\right)\frac{(x-x_{i-1})^2}{2} \right\}_{x_{i-1}}^{x_i}$$

$$= \sum_{i=1}^{n} \left\{ f''_{i-1}\frac{(x_{i-1}-x_i)^4}{24h_i} + f''_i\frac{(x_i-x_{i-1})^4}{24h_i} \right.$$

$$\left. + \left(\frac{f(x_{i-1})}{h_i} - \frac{h_i f''_{i-1}}{6}\right)\frac{(x_{i-1}-x_i)^2}{2} + \left(\frac{f(x_i)}{h_i} - \frac{h_i f''_i}{6}\right)\frac{(x_i-x_{i-1})^2}{2} \right\}$$

Puisque $h_i = x_i - x_{i-1}$, on a:

$$= \sum_{i=1}^{n} \left\{ (f''_{i-1} + f''_i)\frac{h_i^3}{24} + (f(x_{i-1}) + f(x_i))\frac{h_i}{2} - (f''_{i-1} + f''_i)\frac{h_i^3}{12} \right\}$$

$$= \sum_{i=1}^{n} \left\{ (f(x_{i-1}) + f(x_i))\frac{h_i}{2} - (f''_{i-1} + f''_i)\frac{h_i^3}{24} \right\}$$

On obtient ainsi l'approximation suivante de l'intégrale de $f(x)$ dans l'intervalle $[a, b] = [x_0, x_n]$:

$$\int_a^b f(x)dx \simeq \sum_{i=1}^{n} \left\{ \frac{h_i}{2}(f(x_{i-1}) + f(x_i)) - \frac{h_i^3}{24}(f''_{i-1} + f''_i) \right\} \tag{6.40}$$

Dans le cas où les abscisses x_i sont équidistantes ($h_i = h$), on peut simplifier davantage l'expression précédente pour obtenir:

$$\begin{aligned} \int_a^b f(x)dx \quad &\simeq \quad \frac{h}{2}(f(x_0) + 2(f(x_1) + f(x_2) + \cdots + f(x_{n-1})) + f(x_n)) \\ &- \frac{h^3}{24}(f''_0 + 2(f''_1 + f''_2 + \cdots + f''_{n-1}) + f''_n) \end{aligned}$$

Puisque $f''_0 = f''_n = 0$ dans le cas de la spline naturelle, on a:

$$\begin{aligned} \int_a^b f(x)dx \quad &\simeq \quad \frac{h}{2}(f(x_0) + 2(f(x_1) + f(x_2) + \cdots + f(x_{n-1}) + f(x_n)) \\ &- \frac{h^3}{12}(f''_1 + f''_2 + \cdots + f''_{n-1}) \end{aligned} \tag{6.41}$$

Ce résultat mérite quelques commentaires. En effet, cette approximation de l'intégrale comporte deux termes. Le premier terme n'est autre que l'expression de la méthode des trapèzes composée. Le deuxième terme est une approximation du terme d'erreur lié à cette même méthode. En effet, l'erreur d'approximation de la méthode du *trapèze simple* est donnée pour chaque intervalle $[x_{i-1}, x_i]$ par:

$$-\frac{f''(\eta_i)}{12}h^3 \quad \text{où} \quad \eta_i \in [x_{i-1}, x_i]$$

La position exacte du point η_i étant inconnue, on lui attribue la valeur de x_i.

Remarque 6.22
La méthode du trapèze utilise un polynôme de degré 1 dans $[x_{i-1}, x_i]$. On peut interpréter le deuxième terme de droite de l'équation 6.41 comme étant une correction due à la courbure de la fonction $f(x)$ dans cet intervalle. □

Remarque 6.23
Dans le cas général, on utilise l'expression 6.40 pour faire l'approximation de l'intégrale à l'aide de la spline. Si les abscisses x_i sont équidistantes, on utilise de préférence l'expression 6.41. □

6.5 Applications

6.5.1 Courbe des puissances classées

Reprenons l'exemple de la section 5.8 portant sur la courbe des puissances classées d'une entreprise d'électricité. Cette courbe indique la proportion de l'année où la demande d'électricité atteint ou dépasse un niveau de puissance donné (en gigawatts ou GW). L'aire sous cette courbe est tout simplement l'énergie totale E vendue au cours de l'année. Cette donnée est donc importante pour l'entreprise en question. On a en main les données suivantes relatives à une certaine année de référence.

Proportion de l'année	Puissance (GW)
0,0	30
0,1	29
0,2	24
0,5	19
0,8	18
0,9	15
1,0	0

Pour obtenir l'aire sous la courbe et donc l'énergie, il suffit d'intégrer dans l'intervalle $[0, 1]$. Comme les abscisses ne sont pas également distancées, il faut être prudent quant au choix de la méthode d'intégration. Pour obtenir la plus grande précision possible, il est souhaitable de diviser l'intervalle d'intégration en 3 parties et d'utiliser la méthode de Simpson 1/3. Les

3 sous-intervalles sont [0 , 0,2] où $h = 0{,}1$, [0,2 , 0,8] où $h = 0{,}3$ et [0,8 , 1,0] où $h = 0{,}1$. On a alors:

$$\begin{aligned} E &= \frac{0{,}1}{3}(40 + 4 \times 29 + 24) \\ &+ \frac{0{,}3}{3}(24 + 4 \times 19 + 18) + \frac{0{,}1}{3}(18 + 4 \times 15 + 0) \\ &= 20{,}0666 \text{ gigawatts-années} = 175{,}784 \text{ gigawatts-heures} \\ &= 6{,}33 \times 10^{8} \text{ gigawatts-secondes} \\ &= 6{,}33 \times 10^{17} \text{ joules} \\ &= 633 \text{ pétajoules} \end{aligned}$$

6.5.2 Puissance électrique d'un ordinateur

Un ordinateur personnel muni d'un processeur de type 486 est alimenté à une tension $e(t)$ de 120 V et à une fréquence f de 60 Hz. Plus précisément:

$$e(t) = \sqrt{2}\, 120 \cos(2\pi f t)$$

que représente la figure 6.9 (voir IEEE, réf. [14]). Le courant $i(t)$ est ensuite mesuré par un analyseur de puissance qui prend typiquement 256 mesures par période $T = 1/f = 0{,}0166$ s. La figure 6.10 illustre ces mesures. Malgré les petites oscillations, qui sont dues aux erreurs de mesure, on a une assez bonne représentation du signal.

La puissance P (en watts ou W) tirée du réseau est alors définie par:

$$P = \frac{1}{T} \int_0^T e(t) i(t) dt$$

qui est une valeur très importante à connaître en pratique. Pour évaluer P, on utilise une méthode de *Simpson 1/3* composée avec 256 sous-intervalles, c'est-à-dire:

$$h = \frac{T}{256} = 0{,}65 \times 10^{-4}$$

On obtient ainsi $P = 119{,}90$ W. Dans ce cas particulier, il n'est pas nécessaire de chercher à obtenir une précision extrême de la valeur de P, car les données mesurées sont elles-mêmes imprécises.

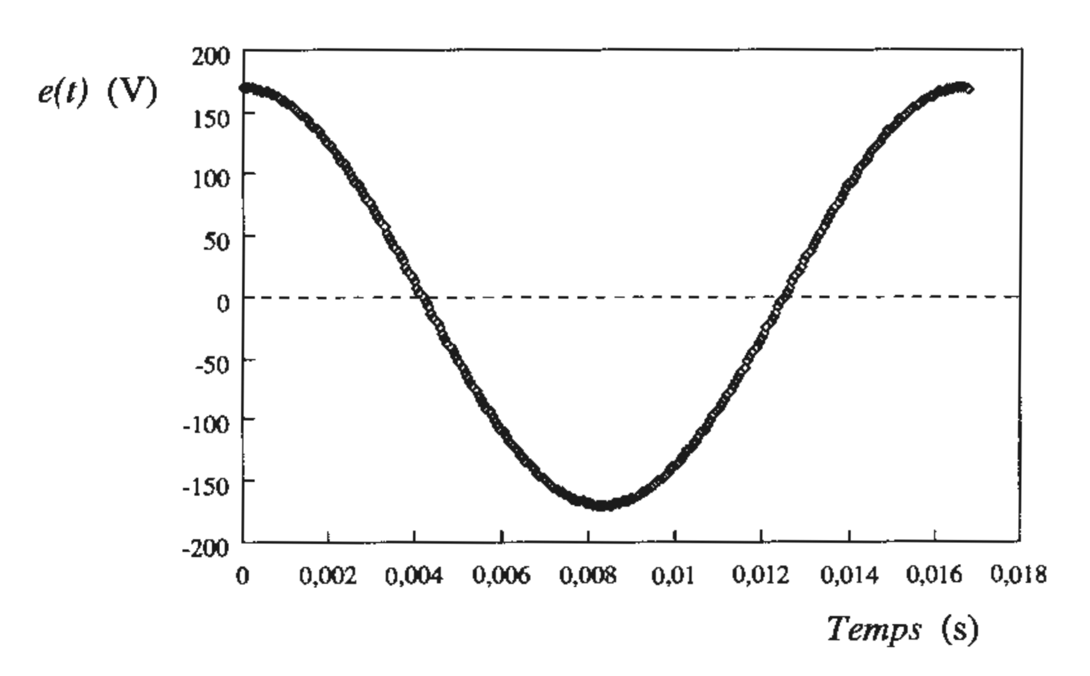


Figure 6.9: Tension $e(t)$

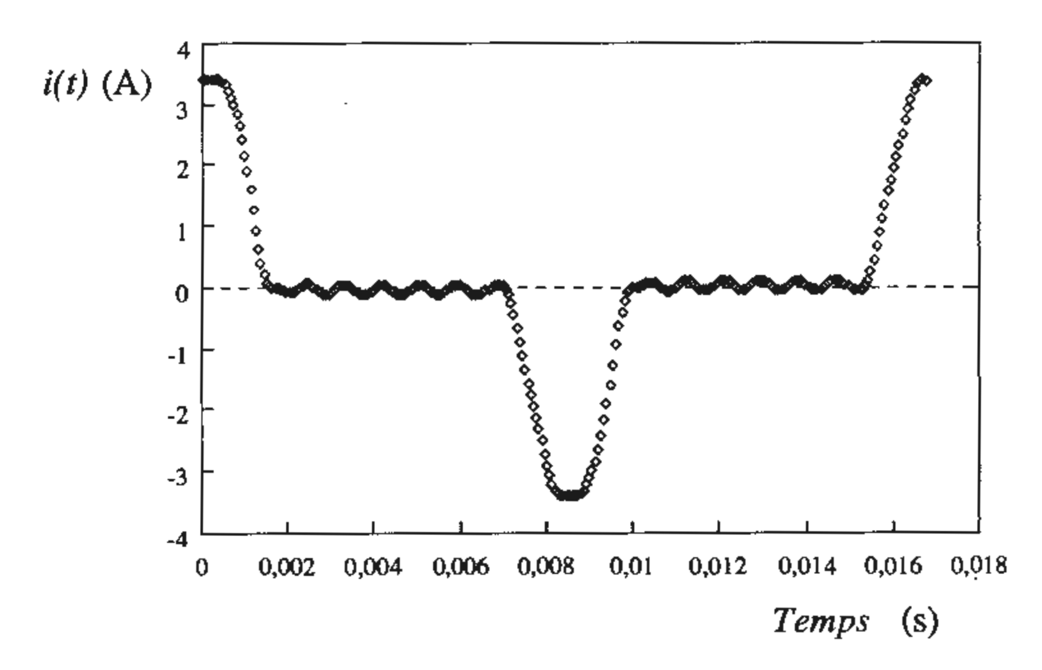


Figure 6.10: Courant mesuré $i(t)$

6.6 Exercices

1. À partir du polynôme de degré 2 passant par les points $(x_0, f(x_0))$, $(x_1, f(x_1))$ et $(x_2, f(x_2))$, obtenir les formules aux différences avant, centrée et arrière d'ordre 2 pour le calcul de $f'(x)$. Déduire également les termes d'erreur.

2. À partir de la relation 6.5, obtenir les termes d'erreur pour les différences d'ordre 2 obtenues à l'exercice précédent.

3. Évaluer la dérivée de $f(x) = e^x$ en $x = 0$ à l'aide des différences avant et arrière d'ordre 2. Prendre $h = 0{,}05$ et $h = 0{,}025$, et calculer le rapport des erreurs commises. Obtenir une approximation encore plus précise de $f'(0)$ à l'aide de l'extrapolation de Richardson.

4. À l'aide des développements de Taylor appropriés, obtenir l'expression du terme d'erreur lié à la différence centrée d'ordre 2 permettant de faire l'approximation de $f'(x)$ (voir l'équation 6.11). Montrer que ce terme d'erreur ne fait intervenir que les puissances paires de h. Que conclure au sujet de l'extrapolation de Richardson dans ce cas?

5. Obtenir la formule de différence centrée d'ordre 4 (voir l'équation 6.13) permettant de faire l'approximation de $f''(x)$. Bien identifier le polynôme d'interpolation qui est nécessaire.

6. Intégrer les polynômes de degré 1, 2 et 3 respectivement permettant d'obtenir les formules du trapèze simple, de Simpson 1/3 simple et de Simpson 3/8 simple. Bien préciser l'intervalle sur lequel porte l'intégration et utiliser le changement de variable $s = (x - x_0)/h$.

7. Intégrer la fonction $f(x) = e^x$ dans l'intervalle $[0, 1]$ en utilisant la méthode des trapèzes composée avec 4 puis avec 8 intervalles. Utiliser l'extrapolation de Richardson avec les deux valeurs obtenues pour obtenir une meilleure approximation. Quel est l'ordre de cette approximation? Comparer les résultats avec la valeur exacte.

8. Refaire le même exercice en utilisant cette fois la méthode de Simpson 1/3 composée.

9. Utiliser la méthode de Simpson 3/8 avec 6 intervalles pour évaluer:

$$\int_1^9 \sqrt{x}\,dx$$

Comparer le résultat avec la valeur exacte.

10. Utiliser la méthode de Boole avec 8 intervalles pour évaluer:

$$\int_0^{\frac{\pi}{4}} \sec x dx$$

Comparer le résultat avec la valeur exacte.

11. Soit la fonction suivante.

x	$f(x)$
0,00	1,570 796 327
0,25	1,318 116 072
0,50	1,047 197 551
0,75	0,722 734 248
1,00	0,000 000 000

Évaluer:

$$\int_0^1 f(x)dx$$

à l'aide de la méthode de Romberg.

12. Donner les expressions complètes de $T_{1,1}$ et de $T_{1,2}$ dans l'intervalle $[a, b]$. Montrer que l'extrapolation de Richardson:

$$\frac{4T_{1,2} - T_{1,1}}{3}$$

donne le même résultat que la méthode de Simpson 1/3 simple. (En général, l'extrapolation de Richardson:

$$\frac{4T_{1,i+1} - T_{1,i}}{3}$$

donne le même résultat que la méthode Simpson 1/3 avec 2^i intervalles).

13. On considère l'intégrale:

$$\int_0^1 \frac{dx}{1+x}$$

a) Donner la valeur exacte de cette intégrale.

b) Calculer les valeurs de $T_{1,i}$ jusqu'à ce que:

$$\frac{|T_{1,i+1} - T_{1,i}|}{|T_{1,i+1}|} < 0{,}005$$

c) À partir des résultats obtenus en b), déterminer la valeur de l'intégrale à l'aide de la méthode de Romberg.

d) *Sans faire de calculs supplémentaires*, donner le résultat que l'on obtiendrait à l'aide de la méthode de Simpson 1/3 avec 8 intervalles.

14. Utiliser une méthode numérique pour évaluer:

$$\int_0^2 \ln x dx$$

15. Quelle serait l'erreur d'approximation si on utilisait la quadrature de Gauss à 3 points pour évaluer:

$$\int_0^3 (3x^5 + 7x^2 + x + 1)dx$$

16. On désire développer une nouvelle formule d'intégration numérique dans l'intervalle $[0, 3h]$ de la forme:

$$\int_0^{3h} f(x)dx \simeq af(h) + bf(2h)$$

a) Déterminer les valeurs des constantes a et b de telle sorte que cette quadrature soit exacte dans le cas de tout polynôme de degré inférieur ou égal à 1.

b) Calculer:

$$\int_0^3 \frac{dx}{1+x}$$

à l'aide de cette quadrature.

c) Calculer l'intégrale donnée en b) à l'aide de la formule de Simpson 1/3 simple.

d) Selon l'écart entre les résultats et la valeur exacte, déterminer le nombre de chiffres significatifs des valeurs obtenues en b) et en c). Conclure brièvement.

17. À l'aide d'une certaine méthode d'intégration numérique, on a évalué:

$$I = \int_0^{\frac{\pi}{2}} \sin x dx$$

en utilisant 3 valeurs de h différentes. On a obtenu les résultats suivants.

h	I
0,1	1,001 235
0,2	1,009 872
0,4	1,078 979

Compte tenu de la valeur exacte de I, déduire l'ordre de convergence de la quadrature employée.

18. Soit l'approximation:

$$\int_{x_0}^{x_0+h} f(x)dx \simeq \frac{h}{4}\left(f(x_0) + 3\left(f(x_0 + \frac{2h}{3}\right)\right)$$

a) Obtenir le développement de Taylor de $f\left(x_0 + \frac{2h}{3}\right)$ jusqu'à l'ordre 5 et proposer une nouvelle expression du terme de droite.

b) Obtenir un développement de Taylor d'ordre 5 du terme de gauche.

Suggestion: Poser:

$$f(x) = f(x_0 + (x - x_0))$$

pour effectuer le développement de Taylor. Par la suite, intégrer les premiers termes de ce développement.

c) Soustraire les expressions obtenues en a) et en b) pour obtenir le premier terme de l'erreur. En déduire l'ordre de la méthode proposée.

d) Quel est le degré de précision de cette méthode?

Chapitre 7

Équations différentielles

7.1 Introduction

La résolution numérique des équations différentielles est probablement le domaine de l'analyse numérique où les applications sont les plus nombreuses. Que ce soit en mécanique des fluides, en transfert de chaleur ou en analyse de structures, on aboutit souvent à la résolution d'équations différentielles, de systèmes d'équations différentielles ou plus généralement d'équations aux dérivées partielles.

Le problème du pendule abordé au chapitre 1 trouvera ici une solution numérique qui sera par la suite analysée et comparée à d'autres solutions approximatives ou quasi analytiques. Parmi leurs avantages, les méthodes numériques permettent d'étudier des problèmes complexes pour lesquels on ne connaît pas de solution analytique, mais qui sont d'un grand intérêt pratique.

Dans ce chapitre comme dans les précédents, les diverses méthodes de résolution proposées sont d'autant plus précises qu'elles sont d'ordre élevé. Nous amorçons l'exposé par des méthodes relativement simples ayant une interprétation géométrique. Elles nous conduiront progressivement à des méthodes plus complexes telles les méthodes de Runge-Kutta d'ordre 4, qui permettent d'obtenir des résultats d'une grande précision. Nous considérons principalement les équations différentielles avec conditions initiales, mais nous ferons une brève incursion du côté des équations différentielles avec conditions aux limites par le biais des méthodes de tir et de différences finies.

Nous prenons comme point de départ la formulation générale d'une équation différentielle d'ordre 1 avec condition initiale. La tâche consiste à déterminer une fonction $y(t)$ solution de:

$$\begin{aligned} y'(t) &= f(t, y(t)) \\ y(t_0) &= y_0 \end{aligned} \tag{7.1}$$

La variable indépendante t représente très souvent (mais pas toujours) le temps. La variable dépendante est notée y et dépend bien sûr de t. La fonction f est pour le moment une fonction quelconque de deux variables que nous supposons suffisamment différentiable. La condition $y(t_0) = y_0$ est la condition initiale et en quelque sorte l'état de la solution au moment où on commence à s'y intéresser. Il s'agit d'obtenir $y(t)$ pour $t \geq t_0$, si on cherche une solution analytique, ou une approximation de $y(t)$, si on utilise une méthode numérique.

Définition 7.1
L'équation différentielle 7.1 est dite *d'ordre 1*, car seule la dérivée d'ordre 1 de la variable dépendante $y(t)$ est présente. Si des dérivées de $y(t)$ d'ordre 2 apparaissaient dans l'équation différentielle 7.1, on aurait une équation d'ordre 2, et ainsi de suite.

Commençons par présenter quelques exemples d'équations différentielles avec condition initiale.

Exemple 7.1
Soit l'équation différentielle du premier ordre:

$$\begin{aligned} y'(t) &= t \\ y(0) &= 1 \end{aligned} \tag{7.2}$$

Voilà certainement l'un des exemples les plus simples que l'on puisse imaginer. En intégrant de chaque côté, on obtient:

$$\int y'(t)dt = \int tdt$$

c'est-à-dire:

$$y(t) = \frac{t^2}{2} + C$$

où C est une constante. Cette dernière expression est la *solution générale de l'équation différentielle* en ce sens qu'elle satisfait $y'(t) = t$, quelle que soit la constante C. Pour déterminer la constante C, il suffit d'imposer la condition initiale:

$$y(0) = 1 = C$$

La *solution particulière* est alors:

$$y(t) = \frac{t^2}{2} + 1$$

qui vérifie à la fois l'équation différentielle et la condition initiale.

● ● ● ● ●

Exemple 7.2
Soit l'équation différentielle:

$$\begin{aligned} y'(t) &= ty(t) \\ y(1) &= 2 \end{aligned} \tag{7.3}$$

Il ne suffit pas dans ce cas d'intégrer les deux côtés de l'équation pour obtenir la solution. On doit d'abord séparer les variables en écrivant par exemple:

$$\frac{dy}{dt} = t\,y(t)$$

qui devient:

$$\frac{dy}{y} = tdt$$

Les variables étant séparées, on peut maintenant faire l'intégration:

$$\ln y = \frac{t^2}{2} + C$$

ou encore

$$y(t) = Ce^{t^2/2}$$

qui est la solution générale. On obtient la solution particulière en imposant la condition initiale:

$$y(1) = 2 = Ce^{1/2}$$

ce qui signifie que:

$$C = 2e^{-1/2}$$

et donc que la solution particulière est:

$$y(t) = 2e^{(t^2-1)/2}$$

• • • • •

Exemple 7.3
Cet exemple est tiré de Gerald et Wheatly (réf. [12]). Il illustre un cas simple d'équation différentielle n'ayant pas de solution analytique. On considère que le taux de variation $N'(t)$ d'une population $N(t)$ de souris dépend:

- du taux de natalité $aN(t)$, qui dépend lui-même du nombre de femelles fertiles;
- du facteur de décès, qui dépend de la quantité de nourriture disponible au temps t. Des études en laboratoire montrent que l'on peut *modéliser* ce phénomène par la fonction $bN^{1,7}(t)$.

On obtient ainsi l'équation différentielle:

$$N'(t) = aN(t) - bN^{1,7}(t)$$

avec la condition initiale qui est dans ce cas la population initiale $N(0) = N_0$. On note que les variables ne sont pas séparables et que les techniques de résolution d'équations différentielles élémentaires ne permettent pas d'obtenir une solution. La résolution numérique est ici essentielle.

• • • • •

Les ouvrages de Simmons (réf. [20]) et de Derrick et Grossman (réf. [7]) contiennent d'autres exemples d'équations différentielles. Notre propos concerne plutôt les méthodes numériques de résolution de ces équations différentielles. À cet égard, nous suivons l'approche de Burden et Faires (réf. [2]), notamment en ce qui concerne la notion d'erreur de troncature locale qui indique l'ordre de précision de la méthode utilisée.

Avec les outils numériques de résolution d'équations différentielles, il n'est plus possible d'obtenir une solution pour toutes les valeurs de la variable indépendante t. On obtient plutôt une approximation de la solution

analytique seulement à *certaines valeurs de t notées t_i* et distancées d'une valeur $h_i = t_{i+1} - t_i$. Dans la plupart des méthodes présentées, cette distance est constante pour tout i et est notée h. On appelle h le *pas de temps.*

Remarque 7.1
On note $y(t_i)$ la *solution analytique* de l'équation différentielle 7.1 en $t = t_i$. On note y_i la *solution approximative* en $t = t_i$ obtenue à l'aide d'une méthode numérique. □

7.2 Méthode d'Euler

La méthode d'Euler est de loin la méthode la plus simple de résolution numérique d'équations différentielles ordinaires. Elle possède une belle interprétation géométrique et son emploi est facile. Toutefois, elle est relativement peu utilisée en raison de sa faible précision.

Reprenons l'équation différentielle 7.1 et considérons plus attentivement la condition initiale $y(t_0) = y_0$. Le but est maintenant d'obtenir une approximation de la solution en $t = t_1 = t_0 + h$. Avant d'effectuer la première itération, il faut déterminer dans quelle direction on doit avancer à partir du point (t_0, y_0) pour obtenir le point (t_1, y_1), qui est une approximation du point $(t_1, y(t_1))$. Nous n'avons pas l'équation de la courbe $y(t)$, mais nous en connaissons la pente $y'(t)$ en $t = t_0$. En effet, l'équation différentielle assure que:

$$y'(t_0) = f(t_0, y(t_0)) = f(t_0, y_0)$$

On peut donc suivre la droite passant par (t_0, y_0) et de pente $f(t_0, y_0)$. L'équation de cette droite, notée $d_0(t)$, est:

$$d_0(t) = y_0 + f(t_0, y_0)(t - t_0)$$

et est illustrée à la figure 7.1.

En $t = t_1$, on a:

$$d_0(t_1) = y_0 + f(t_0, y_0)(t_1 - t_0) = y_0 + hf(t_0, y_0) = y_1$$

En d'autres termes, $d_0(t_1)$ est proche de la solution analytique $y(t_1)$, c'est-à-dire:

$$y(t_1) \simeq y_1 = d_0(t_1) = y_0 + hf(t_0, y_0)$$

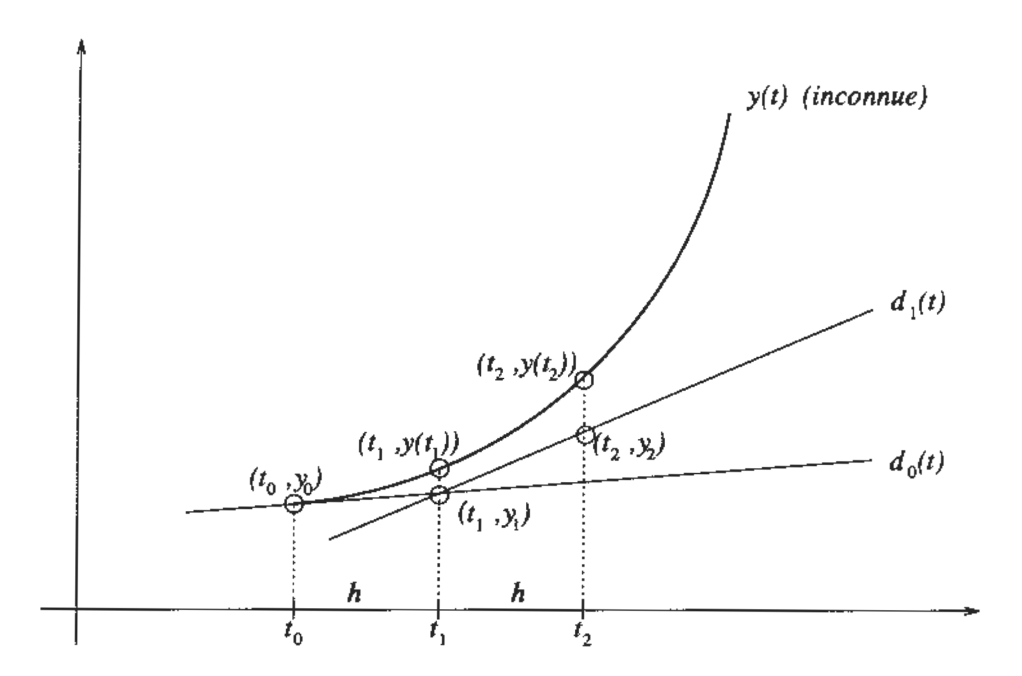


Figure 7.1: Méthode d'Euler

Il est important de noter que, le plus souvent, $y_1 \neq y(t_1)$. Cette inégalité n'a rien pour étonner, mais elle a des conséquences sur la suite du raisonnement. En effet, si on souhaite faire une deuxième itération et obtenir une approximation de $y(t_2)$, on peut refaire l'analyse précédente à partir du point (t_1, y_1). On remarque cependant que la pente de la solution analytique en $t = t_1$ est:

$$y'(t_1) = f(t_1, y(t_1))$$

On ne connaît pas exactement $y(t_1)$, mais nous possédons l'approximation y_1 de $y(t_1)$. On doit alors utiliser l'expression:

$$y'(t_1) = f(t_1, y(t_1)) \simeq f(t_1, y_1)$$

et construire la droite (voir la figure 7.1):

$$d_1(t) = y_1 + f(t_1, y_1)(t - t_1)$$

qui permettra d'estimer $y(t_2)$. *On constate que l'erreur commise à la première itération est réintroduite dans les calculs de la deuxième itération.* On a alors:

$$y(t_2) \simeq y_2 = d_1(t_2) = y_1 + hf(t_1, y_1)$$

Remarque 7.2
Le développement qui précède met en évidence une propriété importante des méthodes numériques de résolution des équations différentielles. En effet, l'erreur introduite à la première itération a des répercussions sur les calculs de la deuxième itération, ce qui signifie que les erreurs se propagent d'une itération à l'autre. Il en résulte de façon générale que l'erreur:

$$|y(t_i) - y_i|$$

augmente légèrement avec i. □

On en arrive donc à l'algorithme suivant.

Algorithme 7.1: Méthode d'Euler

1. Étant donné un pas de temps h, une condition initiale (t_0, y_0) et un nombre maximal d'itérations N

2. Pour $0 \leq n \leq N$:

   $y_{n+1} = y_n + hf(t_n, y_n)$

   $t_{n+1} = t_n + h$

   Écrire t_{n+1} et y_{n+1}

3. Arrêt □

Exemple 7.4
Soit l'équation différentielle (voir Fortin et Pierre, réf. [11]):

$$y'(t) = -y(t) + t + 1$$

et la condition initiale $y(0) = 1$. On a donc $t_0 = 0$ et $y_0 = 1$, et on prend un pas de temps $h = 0{,}1$. De plus, on a:

$$f(t, y) = -y + t + 1$$

On peut donc utiliser la méthode d'Euler et obtenir successivement des approximations de $y(0{,}1), y(0{,}2), y(0{,}3)\cdots$, notées $y_1, y_2, y_3\cdots$. La première itération produit:

$$y_1 = y_0 + hf(t_0, y_0) = 1 + 0{,}1f(0\,,\,1) = 1 + 0{,}1(-1 + 0 + 1) = 1$$

La deuxième itération fonctionne de manière similaire:

$$y_2 = y_1 + hf(t_1, y_1) = 1 + 0{,}1f(0{,}1\,,\,1) = 1 + 0{,}1(-1 + 0{,}1 + 1) = 1{,}01$$

On parvient à:

$$\begin{aligned} y_3 &= y_2 + hf(t_2, y_2) = 1{,}01 + 0{,}1f(0{,}2\,,\,1{,}01) \\ &= 1{,}01 + 0{,}1(-1{,}01 + 0{,}2 + 1) \\ &= 1{,}029 \end{aligned}$$

La tableau suivant rassemble les résultats des dix premières itérations. On peut montrer que la solution analytique de cette équation différentielle est:

$$y(t) = e^{-t} + t$$

ce qui permet de comparer les solutions numérique et analytique, et de constater la croissance de l'erreur. On peut aussi comparer les résultats à la figure 7.2.

t_i	$y(t_i)$	y_i	$\lvert y(t_i) - y_i\rvert$
0,0	1,000 000	1,000 000	0,000 000
0,1	1,004 837	1,000 000	0,004 837
0,2	1,018 731	1,010 000	0,008 731
0,3	1,040 818	1,029 000	0,011 818
0,4	1,070 302	1,056 100	0,014 220
0,5	1,106 531	1,090 490	0,016 041
0,6	1,148 812	1,131 441	0,017 371
0,7	1,196 585	1,178 297	0,018 288
0,8	1,249 329	1,230 467	0,018 862
0,9	1,306 570	1,287 420	0,019 150
1,0	1,367 879	1,348 678	0,019 201

• • • • •

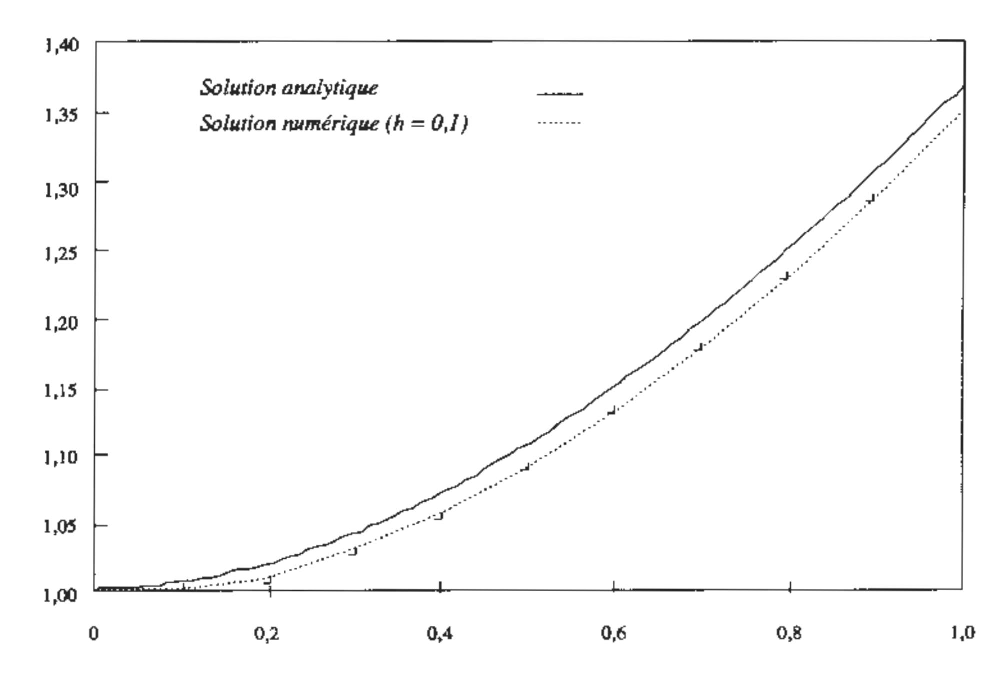


Figure 7.2: Méthode d'Euler: $y'(t) = -y(t) + t + 1$ pour $y(0) = 1$

Les résultats précédents nous amènent à parler de précision et donc d'erreur. La figure 7.2 montre une légère différence entre la solution numérique et la solution analytique. On peut se demander comment se comporte cette erreur en fonction du pas de temps h. La définition qui suit aidera à apporter une réponse. Elle s'applique à la plupart des méthodes étudiées dans ce chapitre.

Définition 7.2
Une méthode de résolution d'équations différentielles est dite *à un pas* si elle est de la forme:

$$y_{n+1} = y_n + h\phi(t_n, y_n) \tag{7.4}$$

où ϕ est une fonction quelconque. Une telle relation est appelée *équation aux différences*. La méthode est à un pas si, pour obtenir la solution en $t = t_{n+1}$, on doit utiliser la solution numérique au temps t_n seulement. On désigne *méthodes à pas multiples* les méthodes qui exigent également la solution numérique aux temps $t_{n-1}, t_{n-2}, t_{n-3} \cdots$.

La méthode d'Euler est bien sûr une méthode à un pas où:

$$\phi(t, y) = f(t, y)$$

Dans ce chapitre, l'attention est principalement portée sur les méthodes à un pas. Nous pouvons maintenant aborder la notion d'erreur de troncature dans le cas de ces méthodes.

Définition 7.3
L'*erreur de troncature locale au point* $t = t_n$ est définie par:

$$\tau_{n+1}(h) = \frac{y(t_{n+1}) - y(t_n)}{h} - \phi(t_n, y(t_n)) \tag{7.5}$$

L'erreur de troncature locale mesure la précision avec laquelle la solution analytique vérifie l'équation aux différences 7.4.

Remarque 7.3
Il est très important de noter que l'on utilise la solution exacte $y(t_n)$ (et non y_n) dans la définition de l'erreur de troncature locale (voir l'équation 7.5). Cela s'explique par le fait que l'on cherche à mesurer l'erreur introduite par l'équation aux différences à un pas donné, en supposant que la méthode était exacte jusque-là. □

Examinons plus avant le cas de la méthode d'Euler ($\phi(t, y) = f(t, y)$). Ici encore, l'outil de travail est le développement de Taylor. En effectuant un développement autour du point $t = t_n$, on trouve:

$$\begin{aligned} y(t_{n+1}) = y(t_n + h) &= y(t_n) + y'(t_n)h + \frac{y''(t_n)h^2}{2} + O(h^3) \\ &= y(t_n) + f(t_n, y(t_n))h + \frac{y''(t_n)h^2}{2} + O(h^3) \end{aligned}$$

puisque $y'(t_n) = f(t_n, y(t_n))$. L'erreur de troncature locale 7.5 devient donc:

$$\tau_{n+1}(h) = \frac{y(t_{n+1}) - y(t_n)}{h} - f(t_n, y(t_n)) = \frac{y''(t_n)h}{2} + O(h^2)$$

ou plus simplement:

$$\tau_{n+1}(h) = O(h)$$

On conclut que l'erreur est d'ordre 1 et qu'elle diminue d'un facteur 2 chaque fois que le pas de temps h est diminué d'un facteur de 2.

Remarque 7.4
Il ne faut pas confondre l'ordre d'une équation différentielle avec l'ordre d'une méthode numérique utilisée pour résoudre cette équation différentielle. □

Exemple 7.5
On tente de résoudre l'équation différentielle:

$$y'(t) = f(t, y) = -y(t) + t + 1$$

en prenant successivement $h = 0{,}1, 0{,}05$ et $0{,}025$. On compare les résultats numériques à la solution analytique à la figure 7.3. On voit nettement diminuer d'un facteur de 2 l'écart entre la solution analytique et la solution numérique chaque fois que le pas de temps h est divisé par 2. Ces résultats confirment que l'erreur de troncature locale de la méthode d'Euler est 1.

● ● ● ● ●

7.3 Méthodes de Taylor

Le développement de Taylor autorise une généralisation immédiate de la méthode d'Euler, qui permet d'obtenir des algorithmes dont l'erreur de troncature locale est d'ordre plus élevé. Nous nous limitons cependant à la méthode de Taylor du second ordre.

On cherche, au temps $t = t_n$, une approximation de la solution en $t = t_{n+1}$. On a immédiatement:

$$\begin{aligned} y(t_{n+1}) &= y(t_n + h) \\ &= y(t_n) + y'(t_n)h + \frac{y''(t_n)h^2}{2} + O(h^3) \end{aligned}$$

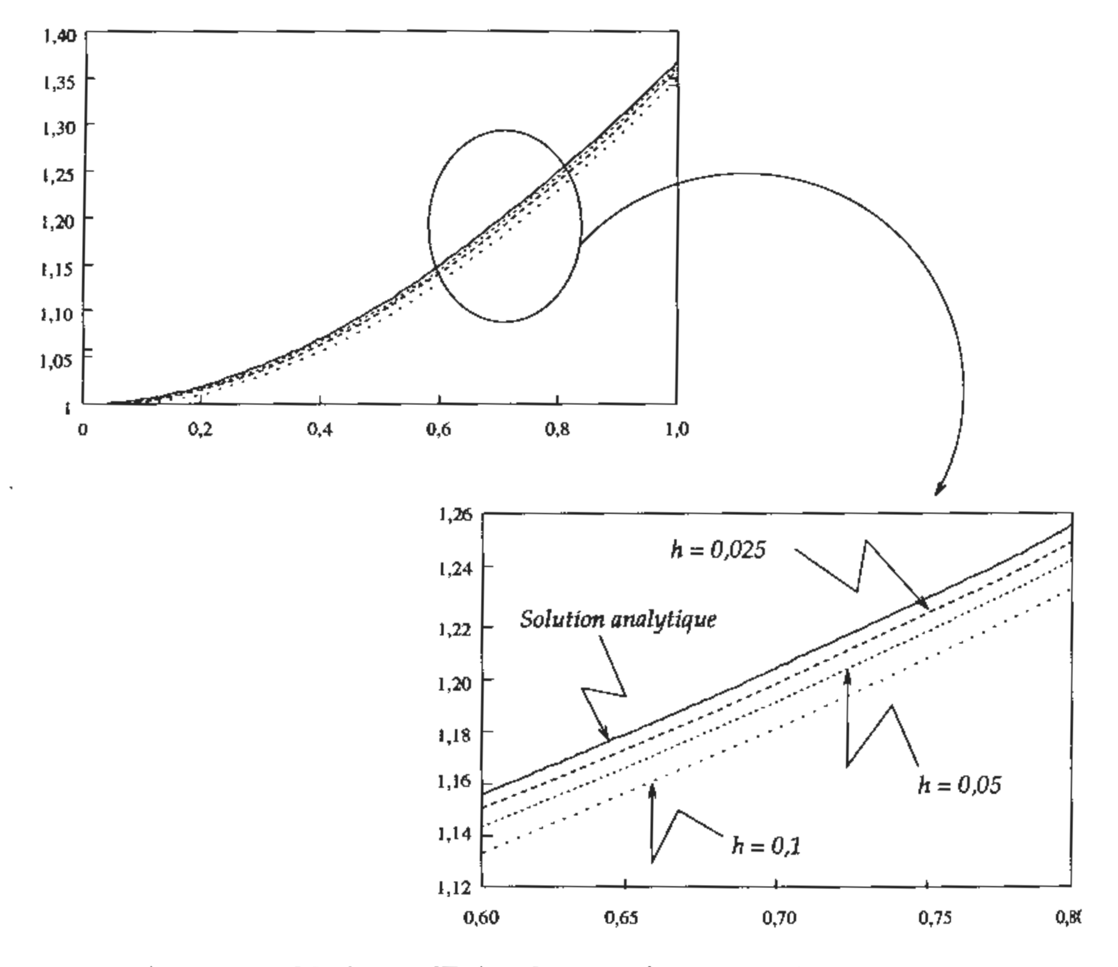


Figure 7.3: Méthode d'Euler: $h = 0{,}1$, $h = 0{,}05$ et $h = 0{,}025$

En se servant de l'équation différentielle 7.1, on trouve:

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + f(t_n, y(t_n))h \\ &+ \frac{f'(t_n, y(t_n))h^2}{2} + O(h^3) \end{aligned}$$

Dans la relation précédente, on voit apparaître la dérivée de la fonction $f(t, y(t))$ par rapport au temps. La règle de dérivation en chaîne (voir Thomas et Finney, réf. [22]) assure que:

$$f'(t, y(t)) = \frac{\partial f(t, y(t))}{\partial t} + \frac{\partial f(t, y(t))}{\partial y} y'(t)$$

c'est-à-dire:

$$f'(t, y(t)) = \frac{\partial f(t, y(t))}{\partial t} + \frac{\partial f(t, y(t))}{\partial y} f(t, y(t))$$

On obtient donc:

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + hf(t_n, y(t_n)) \\ &+ \frac{h^2}{2}\left(\frac{\partial f(t_n, y(t_n))}{\partial t} + \frac{\partial f(t_n, y(t_n))}{\partial y} f(t_n, y(t_n))\right) \\ &+ O(h^3) \end{aligned} \tag{7.6}$$

En négligeant les termes d'ordre supérieur ou égal à 3, on en arrive à poser:

$$\begin{aligned} y(t_{n+1}) &\simeq y(t_n) + hf(t_n, y(t_n)) \\ &+ \frac{h^2}{2}\left(\frac{\partial f(t_n, y(t_n))}{\partial t} + \frac{\partial f(t_n, y(t_n))}{\partial y} f(t_n, y(t_n))\right) \end{aligned} \tag{7.7}$$

qui sera à la base de la méthode de Taylor.

Remarque 7.5

Il importe de préciser l'ordre de troncature locale de cette méthode. Dans ce cas, suivant la notation 7.4, on a:

$$\phi(t, y(t)) = f(t, y(t)) + \frac{h}{2}\left(\frac{\partial f(t, y(t))}{\partial t} + \frac{\partial f(t, y(t))}{\partial y} f(t, y(t))\right)$$

En vertu de la relation 7.6 et de la définition de l'erreur de troncature locale 7.5, il est facile de montrer que:

$$\tau_{n+1}(h) = O(h^2)$$

L'erreur de troncature locale de la méthode de Taylor est d'ordre 2. □

On en arrive donc à l'algorithme suivant.

Algorithme 7.2: Méthode de Taylor d'ordre 2

1. Étant donné un pas de temps h, une condition initiale (t_0, y_0) et un nombre maximal d'itérations N
2. Pour $0 \leq n \leq N$:

$$\begin{aligned} y_{n+1} &= y_n + hf(t_n, y_n) \\ &+ \frac{h^2}{2}\left(\frac{\partial f(t_n, y_n)}{\partial t} + \frac{\partial f(t_n, y_n)}{\partial y} f(t_n, y_n)\right) \qquad (7.8) \\ t_{n+1} &= t_n + h \end{aligned}$$

   Écrire t_{n+1} et y_{n+1}

3. Arrêt □

Remarque 7.6
Dans cet algorithme, on a remplacé la solution analytique $y(t_n)$ par son approximation y_n dans la relation 7.7. On en conclut que les erreurs se propagent d'une itération à une autre. □

Exemple 7.6
Soit l'équation différentielle déjà résolue par la méthode d'Euler:

$$y'(t) = -y(t) + t + 1$$

et la condition initiale $y(0) = 1$. Dans ce cas:

$$f(t, y) = -y + t + 1$$

et

$$\frac{\partial f}{\partial t} = 1 \ \text{ et } \ \frac{\partial f}{\partial y} = -1$$

L'algorithme devient:

$$y_{n+1} = y_n + h(-y_n + t_n + 1) + \frac{h^2}{2}(1 + (-1)(-y_n + t_n + 1))$$

La première itération de la méthode de Taylor d'ordre 2 donne (avec $h = 0{,}1$):

$$y_1 = 1 + 0{,}1(-1 + 0 + 1) + \frac{(0{,}1)^2}{2}(1 + (-1)(-1 + 0 + 1)) = 1{,}005$$

Une deuxième itération donne:

$$\begin{aligned} y_2 &= 1{,}005 + 0{,}1(-1{,}005 + 0{,}1 + 1) + \frac{(0{,}1)^2}{2}(1 + (-1)(-1{,}005 + 0{,}1 + 1)) \\ &= 1{,}019\,025 \end{aligned}$$

Les résultats sont compilés dans le tableau qui suit.

t_i	$y(t_i)$	y_i	$\lvert y(t_i) - y_i \rvert$
0,0	1,000 000	1,000 000	0,000 000
0,1	1,004 837	1,005 000	0,000 163
0,2	1,018 731	1,019 025	0,000 294
0,3	1,040 818	1,041 218	0,000 400
0,4	1,070 302	1,070 802	0,000 482
0,5	1,106 531	1,107 075	0,000 544
0,6	1,148 812	1,149 404	0,000 592
0,7	1,196 585	1,197 210	0,000 625
0,8	1,249 329	1,249 975	0,000 646
0,9	1,306 570	1,307 228	0,000 658
1,0	1,367 879	1,368 541	0,000 662

On remarque que l'erreur est plus petite avec la méthode de Taylor d'ordre 2 qu'avec la méthode d'Euler. Comme on le verra plus loin, cet avantage des méthodes d'ordre plus élevé vaut pour l'ensemble des méthodes de résolution d'équations différentielles.

● ● ● ● ●

Remarque 7.7
Il est possible d'obtenir des méthodes de Taylor encore plus précises en poursuivant le développement de Taylor 7.6 jusqu'à des termes d'ordre élevé. On doit alors évaluer les dérivées de la fonction $f(t, y(t))$ d'ordre de plus en plus élevé, ce qui nécessite le calcul supplémentaire de:

$$\frac{\partial^2 f}{\partial t^2}, \frac{\partial^2 f}{\partial y^2}, \frac{\partial^2 f}{\partial t \partial y}, \cdots, \frac{\partial^{i+j} f}{\partial t^i \partial y^j}$$

Pour cette raison, les méthodes obtenues sont difficiles à utiliser. Il existe cependant un moyen de contourner cette difficulté en développant les méthodes de Runge-Kutta. □

Terminons cette section avec un autre exemple.

Exemple 7.7
Soit l'équation différentielle:

$$\begin{aligned} y'(t) &= ty(t) \\ y(1) &= 2 \end{aligned} \tag{7.9}$$

dont on connaît la solution exacte:

$$y(t) = 2e^{(t^2-1)/2}$$

On a dans ce cas $t_0 = 1, y_0 = 2$ et:

$$f(t, y) = ty$$

d'où on tire que:

$$\frac{\partial f}{\partial t} = y \quad \text{et} \quad \frac{\partial f}{\partial y} = t$$

L'algorithme de la méthode de Taylor devient alors:

$$y_{n+1} = y_n + ht_ny_n + \frac{h^2}{2}(y_n + t_n(t_ny_n))$$

Avec $h = 0{,}5$, la première itération donne:

$$y_1 = 2 + 0{,}5((1)(2)) + 0{,}125\,(2 + 1\,((1)(2))) = 3{,}5$$

La solution analytique est $y(1{,}5) = 3{,}736\,492$. Une deuxième itération donne:

$$y_2 = 3{,}5 + 0{,}5\,((1{,}5)(3{,}5)) + 0{,}125\,(3{,}5 + 1{,}5\,((1{,}5)(3{,}5))) = 7{,}546\,875$$

qui à son tour correspond à la solution analytique $y(2) = 8{,}963\,378$. On constate une erreur assez importante, qui est attribuable à la grande taille du pas de temps h. En effet, si on réduit la taille de h, l'erreur devrait diminuer en $O(h^2)$ car la méthode est d'ordre 2. Cela signifie que, si h est suffisamment petit, la diminution de h par un facteur de 2 réduit l'erreur selon un facteur approximatif de 4. Cependant, lorsqu'on diminue la valeur de h, il faut faire davantage d'itérations pour atteindre le but.

Le tableau qui suit regroupe les approximations de $y(2)$ pour différentes valeurs de h. On y indique également le nombre d'itérations i requis pour obtenir ces approximations. Par exemple, si $h = 0{,}5$, il faut faire deux itérations pour atteindre $t = 2$ à partir de $t = 1$. De même, si $h = 0{,}25$, il faut 4 itérations, et ainsi de suite. On remarque qu'entre chaque ligne du tableau la valeur de h est diminuée selon un facteur de 2, ce qui devrait abaisser l'erreur selon un facteur de 4 puisque la méthode est d'ordre 2. Ce facteur de 4 apparaît lorsque h est suffisamment petit.

h	i	y_i	$\lvert y(2) - y_i \rvert$	Ratio
0,5	2	7,546 875	1,417	—
0,25	4	8,444 292	0,519	2,72
0,125	8	8,804 926	0,158	3,27
0,0625	16	8,919 646	0,0437	3,61
0,031 25	32	8,951 901	0,0114	3,80
0,015 625	64	8,960 439	0,0029	3,87
0,007 812 5	128	8,962 635	0,000 74	3,90

L'avant-dernière colonne du tableau donne l'erreur absolue commise, tandis que la dernière colonne indique le rapport entre l'erreur liée à la valeur de h précédente et celle liée à sa valeur actuelle. On voit bien que l'erreur a tendance à diminuer selon un facteur de 4.

●●●●●

7.4 Méthodes de Runge-Kutta

Il serait avantageux de disposer de méthodes d'ordre de plus en plus élevé *tout en* évitant les désavantages des méthodes de Taylor, qui nécessitent l'évaluation des dérivées partielles de la fonction $f(t, y)$. Une voie est tracée par les méthodes de Runge-Kutta, qui sont calquées sur les méthodes de Taylor du même ordre.

7.4.1 Méthodes de Runge-Kutta d'ordre 2

On a vu que le développement de la méthode de Taylor passe par la relation 7.6:

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + hf(t_n, y(t_n)) \\ &+ \frac{h^2}{2}\left(\frac{\partial f(t_n, y(t_n))}{\partial t} + \frac{\partial f(t_n, y(t_n))}{\partial y} f(t_n, y(t_n))\right) \\ &+ O(h^3) \end{aligned} \tag{7.10}$$

Le but est de remplacer cette dernière relation par une expression équivalente possédant le même ordre de précision ($O(h^3)$). On propose la forme:

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + a_1 hf(t_n, y(t_n)) \\ &+ a_2 hf(t_n + a_3 h, y(t_n) + a_4 h) \end{aligned} \tag{7.11}$$

où on doit déterminer les paramètres a_1, a_2, a_3 et a_4 de telle sorte que les expressions 7.10 et 7.11 aient toutes deux une erreur en $O(h^3)$. On ne trouve par ailleurs aucune dérivée partielle dans cette expression. Pour y arriver, on doit recourir au développement de Taylor en deux variables (voir la section 1.6.2) autour du point $(t_n, y(t_n))$. On a ainsi:

$$\begin{aligned} f(t_n + a_3 h, y(t_n) + a_4 h) &= f(t_n, y(t_n)) + a_3 h \frac{\partial f(t_n, y(t_n))}{\partial t} \\ &+ a_4 h \frac{\partial f(t_n, y(t_n))}{\partial y} + O(h^2) \end{aligned}$$

La relation 7.11 devient alors:

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + (a_1 + a_2)hf(t_n, y(t_n)) \\ &+ a_2a_3h^2\frac{\partial f(t_n, y(t_n))}{\partial t} + a_2a_4h^2\frac{\partial f(t_n, y(t_n))}{\partial y} \\ &+ O(h^3) \end{aligned} \tag{7.12}$$

On voit immédiatement que les expressions 7.10 et 7.12 sont du même ordre. Pour déterminer les coefficients a_i, il suffit de comparer ces deux expressions terme à terme:

- coefficients respectifs de $f(t_n, y(t_n))$:

$$h = (a_1 + a_2)h$$

- coefficients respectifs de $\frac{\partial f(t_n, y(t_n))}{\partial t}$:

$$\frac{h^2}{2} = a_2a_3h^2$$

- coefficients respectifs de $\frac{\partial f(t_n, y(t_n))}{\partial y}$:

$$\frac{h^2}{2}f(t_n, y(t_n)) = a_2a_4h^2$$

On obtient ainsi un système non linéaire de 3 équations comprenant 4 inconnues:

$$\begin{aligned} 1 &= (a_1 + a_2) \\ \frac{1}{2} &= a_2a_3 \\ \frac{f(t_n, y(t_n))}{2} &= a_2a_4 \end{aligned} \tag{7.13}$$

Le système 7.13 est sous-déterminé en ce sens qu'il y a moins d'équations que d'inconnues et qu'il n'a donc pas de solution unique. Cela offre une marge de manœuvre qui favorise la mise au point de plusieurs variantes de la méthode de Runge-Kutta. Voici le choix le plus couramment utilisé.

Méthode d'Euler modifiée

$$a_1 = a_2 = \frac{1}{2}, \quad a_3 = 1 \quad \text{et} \quad a_4 = f(t_n, y(t_n))$$

On établit sans peine que ces coefficients satisfont aux trois équations du système non linéaire. Il suffit ensuite de remplacer ces valeurs dans l'équation 7.11. Pour ce faire, on doit négliger le terme en $O(h^3)$ et remplacer la valeur exacte $y(t_n)$ par son approximation y_n. On obtient alors l'algorithme suivant.

Algorithme 7.3: Méthode d'Euler modifiée

1. Étant donné un pas de temps h, une condition initiale (t_0, y_0) et un nombre maximal d'itérations N

2. Pour $0 \leq n \leq N$:

$$\begin{aligned} \hat{y} &= y_n + hf(t_n, y_n) \\ y_{n+1} &= y_n + \frac{h}{2}\left(f(t_n, y_n) + f(t_{n+1}, \hat{y})\right) \qquad (7.14) \\ t_{n+1} &= t_n + h \end{aligned}$$

   Écrire t_{n+1} et y_{n+1}

3. Arrêt □

Remarque 7.8
Pour faciliter les calculs, l'évaluation de y_{n+1} a été scindée en deux étapes. La variable temporaire $\hat{y}$ correspond tout simplement à une itération de la méthode d'Euler. On fait ainsi une prédiction $\hat{y}$ de la solution en t_{n+1} qui est corrigée (et améliorée) à la deuxième étape de l'algorithme. On parle alors d'une méthode de *prédiction-correction.* □

Exemple 7.8
Soit:

$$y'(t) = -y(t) + t + 1$$

et la condition initiale $y(0) = 1$. On choisit le pas de temps $h = 0{,}1$.

- Itération 1:

  $$\hat{y} = 1 + 0{,}1(-1 + 0 + 1) = 1$$

  qui est le résultat obtenu à l'aide de la méthode d'Euler. La deuxième étape donne:

  $$y_1 = 1 + 0{,}05((-1 + 0 + 1) + (-1 + 0{,}1 + 1)) = 1{,}005$$

- Itération 2:

  De même, la première étape de la deuxième itération donne:

  $$\hat{y} = 1{,}005 + 0{,}1(-1{,}005 + 0{,}1 + 1) = 1{,}0145$$

  La correction conduit à son tour à:

  $$\begin{aligned} y_2 &= 1{,}005 + 0{,}05((-1{,}005 + 0{,}1 + 1) + (-1{,}0145 + 0{,}2 + 1)) \\ &= 1{,}019\,025 \end{aligned}$$

On retrouve ainsi les mêmes résultats qu'avec la méthode de Taylor d'ordre 2. Cette similitude est exceptionnelle et est due au fait que les dérivées partielles d'ordre supérieur on égal à 2 de la fonction $f(t, y)$ sont nulles. On peut montrer dans ce cas particulier que les méthodes de Taylor et d'Euler modifiée sont parfaitement équivalentes. *Ce n'est pas toujours le cas.*

● ● ● ● ●

Une autre méthode de Runge-Kutta d'ordre 2 qui est très utilisée est la *méthode du point milieu*, qui correspond au choix suivant des coefficients a_i.

Méthode du point milieu

$$a_1 = 0, \;\; a_2 = 1, \;\; a_3 = \frac{1}{2} \quad \text{et} \quad a_4 = \frac{f(t_n, y(t_n))}{2}$$

En remplaçant ces valeurs des coefficients a_i dans l'équation 7.11 , on obtient l'algorithme suivant.

Algorithme 7.4: Méthode du point milieu

1. Étant donné un pas de temps h, une condition initiale (t_0, y_0) et un nombre maximal d'itérations N
2. Pour $0 \leq n \leq N$:

$$\begin{aligned} k_1 &= hf(t_n, y_n) \\ y_{n+1} &= y_n + h\left(f(t_n + \frac{h}{2}, y_n + \frac{k_1}{2})\right) \\ t_{n+1} &= t_n + h \end{aligned} \tag{7.15}$$

Écrire t_{n+1} et y_{n+1}

3. Arrêt □

Remarque 7.9
L'algorithme précédent illustre bien pourquoi cette méthode est dite du point milieu. On remarque en effet que la fonction $f(t, y)$ est évaluée au point milieu de l'intervalle $[t_n, t_{n+1}]$. □

Remarque 7.10
Les méthodes d'Euler modifiée et du point milieu étant du même ordre de troncature locale, leur précision est semblable. D'autres choix sont possibles pour les coefficients a_i, mais nous nous limitons aux deux précédents. □

7.4.2 Méthode de Runge-Kutta d'ordre 4

En reprenant le développement de Taylor de la fonction f, mais cette fois jusqu'à l'ordre 5, un raisonnement similaire à celui qui a mené aux méthodes de Runge-Kutta d'ordre 2 aboutit à un système de 8 équations

non linéaires comprenant 10 inconnues (voir Scheid, réf. [25]). Le résultat final est la méthode de Runge-Kutta d'ordre 4, qui représente un outil d'une grande utilité.

Algorithme 7.5: Méthode de Runge-Kutta d'ordre 4

1. Étant donné un pas de temps h, une condition initiale (t_0, y_0) et un nombre maximal d'itérations N
2. Pour $0 \leq n \leq N$:

$$\begin{aligned} k_1 &= hf(t_n, y_n) \\ k_2 &= hf(t_n + \frac{h}{2}, y_n + \frac{k_1}{2}) \\ k_3 &= hf(t_n + \frac{h}{2}, y_n + \frac{k_2}{2}) \\ k_4 &= hf(t_n + h, y_n + k_3) \\ y_{n+1} &= y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ t_{n+1} &= t_n + h \end{aligned} \tag{7.16}$$

   Écrire t_{n+1} et y_{n+1}.

3. Arrêt □

Remarque 7.11
La méthode de Runge-Kutta d'ordre 4 est très fréquemment utilisée en raison de sa grande précision qui est mise en évidence dans l'exemple suivant. □

Exemple 7.9
Soit de nouveau l'équation différentielle:

$$y'(t) = -y(t) + t + 1 \quad (y(0) = 1)$$

Il suffit maintenant d'évaluer les différentes constantes k_i. À la première itération ($h = 0{,}1$), on a:

$$\begin{aligned}
k_1 &= 0{,}1f(0\,,\,1) = 0{,}1(-1+0+1) = 0\\
k_2 &= 0{,}1f(0+0{,}05\,,\,1+0) = 0{,}1(-1+0{,}05+1) = 0{,}005\\
k_3 &= 0{,}1f(0+0{,}05\,,\,1+0{,}0025) = 0{,}1(-1{,}0025+0{,}05+1) = 0{,}004\,75\\
k_4 &= 0{,}1f(0+0{,}1\,,\,1+0{,}004\,75) = 0{,}1(-1{,}004\,75+0{,}1+1) = 0{,}009\,525
\end{aligned}$$

ce qui entraîne que:

$$y_1 = 1 + \frac{1}{6}(0 + 2(0{,}005) + 2(0{,}0045) + 0{,}009\,525) = 1{,}004\,8375$$

Une deuxième itération produit:

$$\begin{aligned}
k_1 &= 0{,}1f(0{,}1\,,\,1{,}004\,8375)\\
&= 0{,}1(-1{,}004\,8375 + 0{,}1 + 1) = 0{,}009\,516\,25\\
\\
k_2 &= 0{,}1f(0{,}15\,,\,1{,}009\,595\,625)\\
&= 0{,}1(-1{,}009\,595\,625 + 0{,}15 + 1) = 0{,}014\,040\,438\\
\\
k_3 &= 0{,}1f(0{,}15\,,\,1{,}011\,857\,719)\\
&= 0{,}1(-1{,}011\,857\,719 + 0{,}15 + 1) = 0{,}013\,814\,2281\\
\\
k_4 &= 0{,}1f(0{,}2\,,\,1{,}018\,651\,728)\\
&= 0{,}1(-1{,}018\,651\,728 + 0{,}2 + 1) = 0{,}018\,134\,8272
\end{aligned}$$

ce qui entraîne que:

$$y_2 = 1{,}004\,8375 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = 1{,}018\,730\,9014$$

Le tableau qui suit compare la solution numérique et la solution exacte, et donne l'erreur absolue.

t_i	$y(t_i)$	y_i	$\|y(t_i) - y_i\|$
0,0	1,0	1,0	0,0
0,1	1,004 837 4180	1,004 837 5000	$0,819 \times 10^{-7}$
0,2	1,018 730 7798	1,018 730 9014	$0,148 \times 10^{-6}$
0,3	1,040 818 2207	1,040 818 4220	$0,210 \times 10^{-6}$
0,4	1,070 320 0460	1,070 320 2889	$0,242 \times 10^{-6}$
0,5	1,106 530 6597	1,106 530 9344	$0,274 \times 10^{-6}$
0,6	1,148 811 6361	1,148 811 9343	$0,298 \times 10^{-6}$
0,7	1,196 585 3034	1,196 585 6186	$0,314 \times 10^{-6}$
0,8	1,249 328 9641	1,249 329 2897	$0,325 \times 10^{-6}$
0,9	1,306 569 6598	1,306 579 9912	$0,331 \times 10^{-6}$
1,0	1,367 879 4412	1,367 879 7744	$0,333 \times 10^{-6}$

On constate que l'erreur se situe autour de 10^{-6}, ce qui se compare avantageusement avec les erreurs obtenues à l'aide de méthodes d'ordre moins élevé. On remarque également une légère croissance de l'erreur au fil des itérations, ce qui indique encore une fois une propagation de l'erreur d'une itération à une autre.

● ● ● ● ●

Il est intéressant de comparer sur une base aussi rigoureuse que possible les différentes méthodes vues jusqu'à maintenant. On a constaté que plus l'ordre d'une méthode est élevé, plus cette méthode est précise. Par contre, plus l'ordre de la méthode est élevé, plus elle est coûteuse en temps de calcul. Par exemple, la méthode d'Euler (d'ordre 1) ne nécessite qu'une seule évaluation de la fonction $f(t, y)$ à chaque pas de temps, alors que la méthode d'Euler modifiée (d'ordre 2) en demande 2 et que la méthode de Runge-Kutta d'ordre 4 exige 4 évaluations de la même fonction. En d'autres termes, la méthode de Runge-Kutta demande à peu près deux fois plus de calculs que la méthode d'Euler modifiée et quatre fois plus que la méthode d'Euler.

Il est raisonnable de se demander s'il n'est pas préférable d'utiliser la méthode d'Euler avec un pas de temps 4 fois plus petit ou la méthode d'Euler modifiée d'ordre 2 avec un pas de temps 2 fois plus petit, plutôt que de se servir de la méthode de Runge-Kutta d'ordre 4. L'exemple qui suit permet de comparer les différentes méthodes sur une base plus équitable.

Exemple 7.10
On considère l'équation différentielle habituelle:

$$y'(t) = -y(t) + t + 1 \quad (y(0) = 1)$$

On recourt à 3 méthodes de résolution: la méthode d'Euler avec un pas $h = 0{,}025$, la méthode d'Euler modifiée avec $h = 0{,}05$ et la méthode de Runge-Kutta d'ordre 4 avec $h = 0{,}1$. Ces valeurs de h permettent de comparer ces 3 méthodes sur la base de coûts de calculs à peu près équivalents. Le tableau suivant présente les résultats obtenus en $t = 1$ pour ces différents choix. La valeur exacte de la solution est $y(1) = 1{,}367\,879\,4412$.

Méthode	h	Nombre de pas	Résultat	Erreur
Euler	0,025	40	1,363 232 374 17	$0{,}464 \times 10^{-2}$
Euler modifiée	0,05	20	1,368 038 621 67	$0{,}159 \times 10^{-3}$
Runge-Kutta	0,1	10	1,367 879 774 41	$0{,}333 \times 10^{-6}$

Les résultats sont éloquents. Même en prenant un pas de temps quatre fois plus petit, la méthode d'Euler reste très imprécise par rapport à celle de Runge-Kutta d'ordre 4. On peut porter le même jugement à la méthode d'Euler modifiée. ***Il est donc généralement préférable d'utiliser des méthodes d'ordre aussi élevé que possible.***

● ● ● ● ●

7.5 Méthodes à pas multiples

Il existe une autre approche de résolution des équations différentielles qui a donné naissance à une famille de méthodes dites à pas multiples. Le principe à la base de ces méthodes consiste à intégrer l'équation différentielle:

$$y'(t) = f(t, y(t))$$

dans l'intervalle $[t_n, t_{n+1}]$, c'est-à-dire:

$$\int_{t_n}^{t_{n+1}} y'(u)du = \int_{t_n}^{t_{n+1}} f(u, y(u))du$$

ou encore

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(u, y(u))du$$

Cela nous amène à un algorithme de la forme:

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f(u, y(u))du \tag{7.17}$$

Il reste à trouver une approximation de l'intégrale présente dans le membre de droite. Pour y arriver, on utilise une interpolation de la fonction $f(u, y(u))$ à partir des valeurs de $y(u)$ calculées aux itérations précédentes. On note $f_n = f(t_n, y_n)$ l'approximation de $f(t_n, y(t_n))$. Il est alors possible de construire une table de différences divisées pour cette fonction et d'effectuer l'interpolation par la méthode de Newton (voir l'équation 5.6). Une première approche consiste à utiliser la table de différences divisées suivante.

t_n	f_n			
		$f[t_n, t_{n-1}]$		
t_{n-1}	f_{n-1}		$f[t_n, t_{n-1}, t_{n-2}]$	
		$f[t_{n-1}, t_{n-2}]$		$f[t_n, t_{n-1}, t_{n-2}, t_{n-3}]$
t_{n-2}	f_{n-2}		$f[t_{n-1}, t_{n-2}, t_{n-3}]$	
		$f[t_{n-2}, t_{n-3}]$		
t_{n-3}	f_{n-3}			

On peut au besoin prolonger cette table. Le polynôme d'interpolation s'écrit:

$$\begin{aligned} p_n(t) &= f_n + f[t_n, t_{n-1}](t - t_n) + f[t_n, t_{n-1}, t_{n-2}](t - t_n)(t - t_{n-1}) \\ &\quad + f[t_n, t_{n-1}, t_{n-2}, t_{n-3}](t - t_n)(t - t_{n-1})(t - t_{n-2}) + \cdots \end{aligned} \tag{7.18}$$

On peut évaluer la fonction $f(t, y(t))$ au moyen de ce polynôme dans l'intervalle $[t_n, t_{n+1}]$. L'évaluation de $p_n(t)$ ne requiert que des valeurs connues provenant des pas de temps antérieurs.

On peut aussi utiliser la table de différences divisées suivante.

t_{n+1}	f_{n+1}			
		$f[t_{n+1}, t_n]$		
t_n	f_n		$f[t_{n+1}, t_n, t_{n-1}]$	
		$f[t_n, t_{n-1}]$		$f[t_{n+1}, t_n, t_{n-1}, t_{n-2}]$
t_{n-1}	f_{n-1}		$f[t_n, t_{n-1}, t_{n-2}]$	
		$f[t_{n-1}, t_{n-2}]$		
t_{n-2}	f_{n-2}			

Le polynôme correspondant est:

$$\begin{aligned} p_n^*(t) &= f_{n+1} + f[t_{n+1}, t_n](t - t_{n+1}) + f[t_{n+1}, t_n, t_{n-1}](t - t_{n+1})(t - t_n) \\ &\quad + f[t_{n+1}, t_n, t_{n-1}, t_{n-2}](t - t_{n+1})(t - t_n)(t - t_{n-1}) + \cdots \end{aligned} \tag{7.19}$$

On remarque immédiatement que l'évaluation de $p_n^*(t)$ requiert la connaissance préalable de $f_{n+1} = f(t_{n+1}, y_{n+1})$. Or, on ne connaît pas encore y_{n+1}. Nous verrons plus loin comment contourner cette difficulté.

Considérons d'abord le polynôme $p_n(t)$. En augmentant successivement le degré du polynôme, on obtient des approximations de plus en plus précises que l'on peut insérer dans la relation 7.17. Par exemple, si on utilise le polynôme de degré 0, on a l'approximation:

$$f(t, y(t)) \simeq p_0(t) = f_n$$

et on trouve:

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f_n \, du = y_n + (t_{n+1} - t_n) f_n = y_n + h f(t_n, y_n)$$

qui n'est rien d'autre que l'expression de la méthode d'Euler. En utilisant maintenant un polynôme de degré 1, on a l'approximation:

$$f(t, y(t)) \simeq p_1(t) = f_n + f[t_n, t_{n-1}](t - t_n)$$

En insérant cette expression dans l'équation 7.17, on obtient:

$$\begin{aligned} y_{n+1} &= y_n + \int_{t_n}^{t_{n+1}} (f_n + f[t_n, t_{n-1}](u - t_n))du \\ &= y_n + (t_{n+1} - t_n)f_n + \frac{(f_n - f_{n-1})}{(t_n - t_{n-1})} \frac{(t_{n+1} - t_n)^2}{2} \\ &= y_n + \frac{h}{2}(3f_n - f_{n-1}) \end{aligned}$$

ou encore

$$y_{n+1} = y_n + \frac{h}{2}(3f(t_n, y_n) + f(t_{n-1}, y_{n-1}))$$

Dans l'équation précédente, on a posé $h = t_n - t_{n-1}$, ce qui suppose que le pas de temps est constant.

On remarque qu'il s'agit d'une *méthode à deux pas* en ce sens que pour obtenir y_{n+1} on doit utiliser y_n et y_{n-1}. Les méthodes vues jusqu'à maintenant (Euler, Taylor, Runge-Kutta, etc.) étaient à un pas. On pourrait continuer ainsi en utilisant des polynômes de degré 2, 3, etc. En substituant ces polynômes dans l'équation 7.17, on obtient les formules d'Adams-Bashforth.

Formules d'Adams-Bashforth

$$\begin{aligned} y_{n+1} &= y_n + hf_n \quad \text{(ordre 1)} \\ y_{n+1} &= y_n + \frac{h}{2}(3f_n - f_{n-1}) \quad \text{(ordre 2)} \\ y_{n+1} &= y_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2}) \quad \text{(ordre 3)} \\ y_{n+1} &= y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) \quad \text{(ordre 4)} \end{aligned}$$

Remarque 7.12

On définit l'erreur de troncature locale liée aux méthodes à pas multiples d'une manière semblable à ce que l'on fait dans le cas des méthodes à un

pas. Dans ce qui suit, on indique sans démonstration l'ordre de l'erreur de troncature locale au fur et à mesure des besoins. On constate qu'en utilisant un polynôme de degré n dans la relation 7.17 on obtient une méthode à $(n+1)$ pas dont l'erreur de troncature locale est d'ordre $(n+1)$. □

Passons maintenant au polynôme $p_n^*(t)$. On peut reprendre le raisonnement précédent, mais cette fois-ci en prenant l'approximation:

$$f(t, y(t)) \simeq p_n^*(t)$$

En particulier, le polynôme de degré 0 est:

$$p_0^*(t) = f_{n+1}$$

et celui de degré 1 est:

$$p_1^*(t) = f_{n+1} + f[t_{n+1}, t_n](t - t_{n+1})$$

On peut ainsi passer à des polynômes de degré de plus en plus élevé dans l'équation 7.17 et obtenir les formules d'Adams-Moulton, qui sont résumées dans le tableau suivant.

Formules d'Adams-Moulton
$y_{n+1} = y_n + hf_{n+1}$ (ordre 1)
$y_{n+1} = y_n + \frac{h}{2}(f_{n+1} + f_n)$ (ordre 2)
$y_{n+1} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1})$ (ordre 3)
$y_{n+1} = y_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2})$ (ordre 4)

Les formules d'Adams-Moulton sont dites *implicites* en ce sens que les relations qui permettent d'évaluer y_{n+1} dépendent de y_{n+1} lui-même. Pour contourner cette difficulté, on combine les formules d'Adams-Bashforth et

d'Adams-Moulton en des schémas dits *prédicteurs-correcteurs*. Il s'agit simplement d'utiliser les schémas d'Adams-Bashforth pour obtenir une première approximation y^p_{n+1} de y_{n+1}, qui est l'étape de prédiction. On fait appel ensuite aux formules d'Adams-Moulton pour corriger et éventuellement améliorer cette approximation. Il est important de remarquer que, dans ce cas, l'évaluation de f_{n+1} dans les formules d'Adams-Moulton repose sur l'emploi de y^p_{n+1}, c'est-à-dire:

$$f_{n+1} \simeq f^p_{n+1} = f(t_{n+1}, y^p_{n+1})$$

On obtient ainsi les schémas suivants.

Schémas de prédiction-correction
$y^p_{n+1} = y_n + hf_n$ $y_{n+1} = y_n + hf^p_{n+1}$ (ordre 1)
$y^p_{n+1} = y_n + \frac{h}{2}(3f_n - f_{n-1})$ $y_{n+1} = y_n + \frac{h}{2}(f^p_{n+1} + f_n)$ (ordre 2)
$y^p_{n+1} = y_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2})$ $y_{n+1} = y_n + \frac{h}{12}(5f^p_{n+1} + 8f_n - f_{n-1})$ (ordre 3)
$y^p_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})$ $y_{n+1} = y_n + \frac{h}{24}(9f^p_{n+1} + 19f_n - 5f_{n-1} + f_{n-2})$ (ordre 4)

Remarque 7.13

L'initialisation des méthodes de prédiction-correction nécessite l'usage d'une méthode à un pas. Si on prend par exemple le schéma d'ordre 4, il est clair que n doit être plus grand ou égal à 3, car autrement on aurait besoin de y_{-1}, y_{-2}, etc. Or, au départ, seul y_0 est connu, provenant de la condition initiale. Les valeurs de y_1, de y_2 et de y_3 doivent être calculées à l'aide d'une autre méthode. Le plus souvent, on recourt à une méthode de Runge-Kutta qui est au moins du même ordre de convergence que la méthode de prédiction-correction que l'on souhaite utiliser. □

Exemple 7.11

On reprend l'équation différentielle:

$$y'(t) = -y(t) + t + 1 \quad (y(0) = 1)$$

On fait appel cette fois aux méthodes de prédiction-correction d'ordre 2 et 4. Les premiers pas de temps sont calculés par une méthode de Runge-Kutta d'ordre 4 qui a déjà servi à résoudre cette équation différentielle. La méthode de prédiction-correction d'ordre 2 exige de connaître y_0, qui vaut 1, et y_1, qui a été calculé au préalable à l'aide de la méthode de Runge-Kutta d'ordre 4 et qui vaut 1,004 8375 (une méthode de Runge-Kutta d'ordre 2 aurait été suffisante).

La première itération donne d'abord une prédiction:

$$\begin{aligned} y_2^p &= y_1 + \frac{h}{2}(3f(t_1, y_1) - f(t_0, y_0)) \\ &= 1{,}004\,8375 + \frac{0{,}1}{2}(3f(0{,}1\,,\,1{,}004\,8375) - f(0{,}0\,,\,1{,}0)) \\ &= 1{,}004\,8375 + 0{,}05(3(-1{,}004\,8375 + 0{,}1 + 1) - (-1 + 0 + 1)) \\ &= 1{,}019\,111\,875 \end{aligned}$$

et ensuite une correction:

$$\begin{aligned} y_2 &= y_1 + \frac{h}{2}(f(t_2, y_2^p) + f(t_1, y1)) \\ &= 1{,}004\,8375 + \frac{0{,}1}{2}(f(0{,}2\,,\,1{,}019\,111\,875) + f(0{,}1\,,\,1{,}004\,8375)) \\ &= 1{,}004\,8375 + 0{,}05(-1{,}019\,111\,875 + 0{,}2 + 1) + (-1{,}004\,8375 + 0{,}1 + 1)) \\ &= 1{,}018\,640\,031 \end{aligned}$$

Les autres itérations sont résumées dans le tableau suivant.

t	y_n^p	y_n	e_n
0,0	—	1,000 000 000	0
0,1	—	1,004 837 500	$0{,}819\,640\,40 \times 10^{-7}$
0,2	1,019 111 875	1,018 640 031	$0{,}907\,218\,28 \times 10^{-4}$
0,3	1,041 085 901	1,040 653 734	$0{,}164\,486\,07 \times 10^{-3}$
0,4	1,070 487 675	1,070 096 664	$0{,}223\,381\,96 \times 10^{-3}$
0,5	1,106 614 851	1,106 261 088	$0{,}269\,571\,40 \times 10^{-3}$
0,5	1,148 826 758	1,148 506 695	$0{,}304\,940\,11 \times 10^{-3}$
0,7	1,196 543 746	1,196 254 173	$0{,}331\,129\,90 \times 10^{-3}$
0,8	1,249 241 382	1,248 979 396	$0{,}361\,493\,25 \times 10^{-3}$
0,9	1,306 445 195	1,306 208 166	$0{,}367\,978\,57 \times 10^{-3}$
1,0	1,367 725 911	1,367 511 462	$0{,}367\,978\,57 \times 10^{-3}$

L'erreur à $t = 0{,}1$ est beaucoup plus faible qu'aux autres valeurs de t puisque la solution numérique à cet endroit a été calculée à l'aide d'une méthode d'ordre 4. Cela explique également l'absence de prédicteur pour cette valeur. De manière similaire, la méthode de prédiction-correction d'ordre 4 requiert le calcul de y_1, de y_2 et de y_3 à l'aide de la méthode de Runge-Kutta d'ordre 4. Par la suite, il suffit d'utiliser l'algorithme:

$$\begin{aligned} y_{n+1}^p &= y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) \\ y_{n+1} &= y_n + \frac{h}{24}(9f_{n+1}^p + 19f_n - 5f_{n-1} + f_{n-2}) \end{aligned}$$

pour $n \geq 3$. La première itération s'effectue comme suit:

$$\begin{aligned}
y_4^p &= y_3 + \frac{0{,}1}{24}(55f(t_3, y_3) - 59f(t_2, y_2) + 37f(t_1, y_1) - 9f(t_0, y_0)) \\
&= 1{,}040\,8184 + \frac{0{,}1}{24}(55f(0{,}3\,,\, 1{,}040\,8184) - 59f(0{,}2\,,\, 1{,}018\,7309) \\
&\quad + 37f(0{,}1\,,\, 1{,}004\,8375) - 9f(0\,,\, 1)) \\
&= 1{,}107\,0323 \\
y_4 &= y_3 + \frac{0{,}1}{24}(9f(t_4, y_4^p) + 19f(t_3, y_3) - 5f(t_2, y_2) + f(t_1, y_1)) \\
&= 1{,}040\,8184 + \frac{0{,}1}{24}(9f(0{,}4\,,\, 1{,}107\,0323) + 19f(0{,}3\,,\, 1{,}040\,8184) \\
&\quad - 5f(0{,}2\,,\, 1{,}018\,7309) + f(0{,}1\,,\, 1{,}004\,8375)) \\
&= 1{,}070\,3199
\end{aligned}$$

On obtient enfin les résultats suivants:

t	y_n^p	y_n	e_n
0,0	—	1,000 0000	0
0,1	—	1,004 8375	$0{,}819\,640 \times 10^{-7}$
0,2	—	1,018 7309	$0{,}148\,328 \times 10^{-6}$
0,3	—	1,040 8184	$0{,}201\,319 \times 10^{-6}$
0,4	1,107 0323	1,070 3199	$0{,}127\,791 \times 10^{-6}$
0,5	1,106 5332	1,106 5303	$0{,}391\,302 \times 10^{-6}$
0,6	1,148 8136	1,148 8110	$0{,}603\,539 \times 10^{-6}$
0,7	1,196 5869	1,196 5845	$0{,}772\,415 \times 10^{-6}$
0,8	1,249 3302	1,249 3281	$0{,}903\,669 \times 10^{-6}$
0,9	1,306 5706	1,306 5687	$0{,}100\,294 \times 10^{-5}$
1,0	1,367 8801	1,367 8784	$0{,}107\,514 \times 10^{-5}$

L'erreur est ici beaucoup plus petite qu'avec la méthode d'ordre 2.

● ● ● ● ●

Remarque 7.14

La précision des méthodes de prédiction-correction est comparable à celle des méthodes à un pas d'ordre équivalent. Si on considère, par exemple, le schéma

d'ordre 4, les résultats sont comparables à ceux de la méthode de Runge-Kutta d'ordre 4. Par contre, si on calcule le coût lié à l'emploi d'une méthode selon le nombre d'évaluations de la fonction $f(t, y)$ à chaque itération, on se rend compte que la méthode de prédiction-correction d'ordre 4 est moins coûteuse. En effet, chaque itération de la méthode de Runge-Kutta nécessite 4 évaluations de la fonction $f(t, y)$, alors que la méthode de prédiction-correction n'exige que 2 nouvelles évaluations, compte tenu de celles qui ont déjà été effectuées aux itérations précédentes. Cela est également vrai dans le cas des méthodes d'ordre 2 et 3. □

7.6 Systèmes d'équations différentielles

Cette section traite de la façon dont on peut utiliser les méthodes de résolntion d'équations différentielles ordinaires dans le cas de *systèmes d'équations différentielles avec conditions initiales.* Fort heureusement, il suffit d'adapter légèrement les méthodes déjà vues.

La forme générale d'un système de m équations différentielles avec conditions initiales s'écrit:

$$\begin{array}{rcll}
y_1'(t) & = & f_1(t, y_1(t), y_2(t), \cdots, y_m(t)) & (y_1(t_0) = y_{1,0}) \\
y_2'(t) & = & f_2(t, y_1(t), y_2(t), \cdots, y_m(t)) & (y_2(t_0) = y_{2,0}) \\
y_3'(t) & = & f_3(t, y_1(t), y_2(t), \cdots, y_m(t)) & (y_3(t_0) = y_{3,0}) \\
 & & \vdots & \vdots \\
y_m'(t) & = & f_m(t, y_1(t), y_2(t), \cdots, y_m(t)) & (y_m(t_0) = y_{m,0})
\end{array} \tag{7.20}$$

Ici encore, on note $y_i(t_n)$, la valeur exacte de la i^e variable dépendante en $t = t_n$ et $y_{i,n}$, son approximation numérique.

Remarque 7.15
Ces m équations sont *couplées* en ce sens que l'éqnation différentielle régissant la variable dépendante $y_i(t)$ peut dépendre de toutes les autres variables dépendantes. On remarque de plus les m conditions initiales qui assurent l'unicité de la solution sous diverses hypothèses que nous ne précisons pas. □

Parmi les techniques de résolution des systèmes d'équations différentielles, nous ne présentons que la méthode de Runge-Kutta d'ordre 4. Il

est possible d'utiliser également les autres méthodes déjà vues, mais leur précision s'avère souvent insuffisante.

Algorithme 7.6: Méthode de Runge-Kutta d'ordre 4

1. Étant donné un pas de temps h, des conditions initiales $(t_0, y_{1,0}, y_{2,0}, \cdots, y_{m,0}$ et un nombre maximal d'itérations N

2. Pour $0 \leq n \leq N$:

$$\begin{aligned}
& \text{Pour } i = 1, 2, 3, \cdots, m : \\
k_{i,1} &= hf_i(t_n, y_{1,n}, y_{2,n}, \cdots, y_{m,n}) \\
& \text{Pour } i = 1, 2, 3, \cdots, m : \\
k_{i,2} &= hf_i(t_n + \frac{h}{2}, y_{1,n} + \frac{k_{1,1}}{2}, y_{2,n} + \frac{k_{2,1}}{2}, \cdots, y_{m,n} + \frac{k_{m,1}}{2}) \\
& \text{Pour } i = 1, 2, 3, \cdots, m : \\
k_{i,3} &= hf_i(t_n + \frac{h}{2}, y_{1,n} + \frac{k_{1,2}}{2}, y_{2,n} + \frac{k_{2,2}}{2}, \cdots, y_{m,n} + \frac{k_{m,2}}{2}) \\
& \text{Pour } i = 1, 2, 3, \cdots, m : \\
k_{i,4} &= hf_i(t_n + h, y_{1,n} + k_{1,3}, y_{2,n} + k_{2,3}, \cdots, y_{m,n} + k_{m,3}) \\
& \text{Pour } i = 1, 2, 3, \cdots, m : \\
y_{i,n+1} &= y_{i,n} + \frac{1}{6}(k_{i,1} + 2k_{i,2} + 2k_{i,3} + k_{i,4}) \\
t_{n+1} &= t_n + h
\end{aligned} \tag{7.21}$$

Écrire t_{n+1} et $y_{i,n+1}$ pour $i = 1, 2, 3, \cdots, m$

3. Arrêt □

Remarque 7.16
Cet algorithme est complexe en apparence. Pour bien en comprendre le principe, il suffit d'y voir une application de la méthode de Runge-Kutta à chacune des équations différentielles. De plus, *il est nécessaire de calculer les m constantes $k_{i,1}$ avant de passer au calcul des constantes $k_{i,2}$ et ainsi de suite.*
□

Exemple 7.12
Soit le système de deux équations différentielles suivant:

$$\begin{array}{rcll} y_1'(t) & = & y_2(t) & (y_1(0) = 2) \\ y_2'(t) & = & 2y_2(t) - y_1(t) & (y_2(0) = 1) \end{array} \tag{7.22}$$

dont la solution analytique est:

$$\begin{array}{rcl} y_1(t) & = & 2e^t - te^t \\ y_2(t) & = & e^t - te^t \end{array}$$

On a alors:

$$\begin{array}{rcl} f_1(t, y_1(t), y_2(t)) & = & y_2(t) \\ f_2(t, y_1(t), y_2(t)) & = & 2y_2(t) - y_1(t) \end{array}$$

et la condition initiale $(t_0, y_{1,0}, y_{2,0}) = (0\,,\,2\,,\,1)$. Si on prend par exemple $h = 0{,}1$, on trouve:

$$\begin{array}{rcl} k_{1,1} & = & 0{,}1(f_1(0\,,\,2\,,\,1)) = 0{,}1 \\ k_{2,1} & = & 0{,}1(f_2(0\,,\,2\,,\,1)) = 0 \end{array}$$

$$\begin{array}{rcl} k_{1,2} & = & 0{,}1(f_1(0{,}05\,,\,2{,}05\,,\,1{,}0)) = 0{,}1 \\ k_{2,2} & = & 0{,}1(f_2(0{,}05\,,\,2{,}05\,,\,1{,}0)) = -0{,}005 \end{array}$$

$$\begin{array}{rcl} k_{1,3} & = & 0{,}1(f_1(0{,}05\,,\,2{,}05\,,\,0{,}9975)) = 0{,}099\,75 \\ k_{2,3} & = & 0{,}1(f_2(0{,}05\,,\,2{,}05\,,\,0{,}9975)) = -0{,}0055 \end{array}$$

$$\begin{array}{rcl} k_{1,4} & = & 0{,}1(f_1(0{,}1\,,\,2{,}099\,75\,,\,0{,}9945)) = 0{,}099\,45 \\ k_{2,4} & = & 0{,}1(f_2(0{,}1\,,\,2{,}099\,75\,,\,0{,}9945)) = -0{,}011\,075 \end{array}$$

$$\begin{array}{rcl} y_{1,1} & = & y_{1,0} + \frac{1}{6}(0{,}1 + 2(0{,}1) + 2(0{,}099\,75) + 0{,}099\,45) \\ & & 2{,}099\,825 \end{array}$$

$$\begin{array}{rcl} y_{2,1} & = & y_{2,0} + \frac{1}{6}(0 + 2(-0{,}005) + 2(-0{,}0055) + (-0{,}011\,075)) \\ & & 0{,}994\,651\,667 \end{array}$$

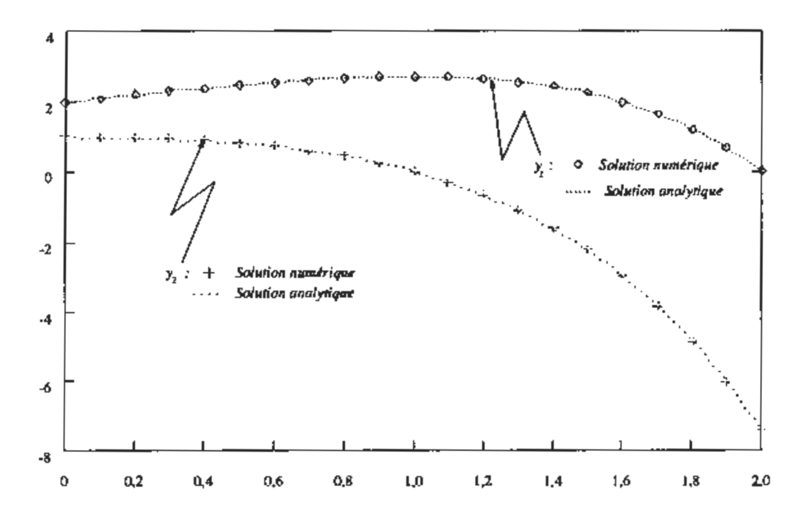


Figure 7.4: $y_1'(t) = y_2(t)$ $(y_1(0) = 2)$ et $y_2'(t) = 2y_2(t) - y_1(t)$ $(y_2(0) = 1)$

La figure 7.4 illustre les solutions analytiques $y_1(t)$ et $y_2(t)$ de même que leurs approximations respectives. On peut y apprécier la grande précision des résultats.

●●●●●

7.7 Équations d'ordre supérieur

Dans la section précédente, nous nous sommes intéressés à la résolution d'un système de m équations différentielles d'ordre 1. Passons maintenant à la résolution numérique d'une équation différentielle d'ordre m avec conditions initiales. Ici encore, nous n'avons pas besoin de développer de nouvelles méthodes numériques, car une équation différentielle d'ordre m avec conditions initiales est parfaitement équivalente à un système de m équations différentielles d'ordre 1.

La forme générale d'une équation différentielle d'ordre m avec conditions initiales est:

$$y^{(m)}(t) = f(t, y(t), y^{(1)}(t), y^{(2)}(t), \cdots, y^{(m-1)}(t)) \tag{7.23}$$

où $y^{(i)}(t)$ désigne la i^{e} dérivée de $y(t)$. Pour assurer l'unicité de la solution,

on ajoute les m conditions initiales:

$$\begin{array}{lll} y(t_0) & = & c_1 \\ y^{(1)}(t_0) & = & c_2 \\ y^{(2)}(t_0) & = & c_3 \\ \vdots & & \vdots \\ y^{(m-2)}(t_0) & = & c_{m-1} \\ y^{(m-1)}(t_0) & = & c_m \end{array} \tag{7.24}$$

Ces conditions portent sur la fonction $y(t)$ et ses $(m-1)$ premières dérivées à $t = t_0$.

Remarque 7.17
La première dérivée de la variable $y(t)$ est notée $y'(t)$ ou $y^{(1)}(t)$ selon la situation. On se doit également de distinguer la dérivée seconde $y^{(2)}(t)$ du carré de la fonction $y(t)$, qui est noté $(y(t))^2$. □

Théorème 7.1
L'équation différentielle d'ordre m 7.23 avec les m conditions initiales 7.24 est équivalente au système de m équations d'ordre 1 suivant:

$$\begin{array}{lllllll} y_1'(t) & = & y_2(t) & & y_1(t_0) & = & c_1 \\ y_2'(t) & = & y_3(t) & & y_2(t_0) & = & c_2 \\ y_3'(t) & = & y_4(t) & & y_3(t_0) & = & c_3 \\ \vdots & & \vdots & & \vdots & & \vdots \\ y_{m-1}'(t) & = & y_m(t) & & y_{m-1}(t_0) & = & c_{m-1} \\ y_m'(t) & = & f(t, y_1(t), y_2(t), \cdots, y_m(t)) & & y_m(t_0) & = & c_m \end{array} \tag{7.25}$$

Démonstration:
Pour voir l'équivalence, il suffit d'introduire les m variables suivantes:

$$\begin{array}{lll} y_1(t) & = & y(t) \\ y_2(t) & = & y^{(1)}(t) \\ y_3(t) & = & y^{(2)}(t) \\ \vdots & & \vdots \\ y_{m-1}(t) & = & y^{(m-2)}(t) \\ y_m(t) & = & y^{(m-1)}(t) \end{array} \tag{7.26}$$

On a alors:

$$\begin{array}{lllll} y_1'(t) & = & y^{(1)}(t) & = & y_2(t) \\ y_2'(t) & = & y^{(2)}(t) & = & y_3(t) \\ y_3'(t) & = & y^{(3)}(t) & = & y_4(t) \\ \vdots & & \vdots & & \vdots \\ y_{m-1}'(t) & = & y^{(m-1)}(t) & = & y_m(t) \\ y_m'(t) & = & y^{(m)}(t) & = & f(t, y_1(t), y_2(t), \cdots, y_m(t)) \end{array} \quad (7.27)$$

Un raisonnement similaire permet de déterminer les conditions initiales associées aux variables $y_i(t)$:

$$\begin{array}{lllll} y_1(t_0) & = & y(t_0) & = & c_1 \\ y_2(t_0) & = & y^{(1)}(t_0) & = & c_2 \\ y_3(t_0) & = & y^{(2)}(t_0) & = & c_3 \\ \vdots & & \vdots & & \vdots \\ y_{m-1}(t_0) & = & y^{(m-2)}(t_0) & = & c_{m-1} \\ y_m(t_0) & = & y^{(m-1)}(t_0) & = & c_m \end{array} \quad (7.28)$$

Il est alors clair que le système 7.27 sous les conditions initiales 7.28 est un cas particulier de systèmes d'équations d'ordre 1 dont la forme générale est donnée par l'équation 7.20. □

Exemple 7.13
Soit l'équation différentielle d'ordre 2:

$$y^{(2)}(t) = -y^{(1)}(t) + (y(t))^2 + t^2 - 5$$

avec les conditions initiales $y(0) = 1$ et $y^{(1)}(0) = 2$. Suivant la démarche décrite, on pose:

$$\begin{array}{lll} y_1(t) & = & y(t) \\ y_2(t) & = & y^{(1)}(t) \end{array}$$

On obtient alors le système de 2 équations différentielles du premier ordre suivant:

$$\begin{array}{llll} y_1'(t) & = & y_2(t) & (y_1(0) = 1) \\ y_2'(t) & = & -y_2(t) + (y_1(t))^2 + t^2 - 5 & (y_2(0) = 2) \end{array}$$

● ● ● ● ●

Exemple 7.14
Soit l'équation différentielle du troisième ordre:

$$y^{(3)}(t) = (y^{(2)}(t))^2 + 2y^{(1)}(t) + (y(t))^3 + t^4 + 1$$

avec les conditions initiales $y(1) = 1$, $y^{(1)}(1) = 0$ et $y^{(2)}(1) = 3$. On pose:

$$\begin{array}{rcl} y_1(t) & = & y(t) \\ y_2(t) & = & y^{(1)}(t) \\ y_3(t) & = & y^{(2)}(t) \end{array}$$

On obtient le système de 3 équations différentielles du premier ordre suivant:

$$\begin{array}{rcll} y_1'(t) & = & y_2(t) & (y_1(1) = 1) \\ y_2'(t) & = & y_3(t) & (y_2(1) = 0) \\ y_3'(t) & = & (y_3(t))^2 + 2y_2(t) + (y_1(t))^3 + t^4 + 1 & (y_3(1) = 3) \end{array}$$

● ● ● ● ●

Remarque 7.18
Une fois l'équation d'ordre m transformée en un système de m équations différentielles d'ordre 1, on peut recourir à l'algorithme 7.21 pour sa résolution.
□

7.8 Méthode de tir

Dans cette section, nous faisons l'étude des équations différentielles linéaires d'ordre 2 avec conditions aux limites de la forme:

$$\begin{array}{c} y''(x) = a_2(x)y'(x) + a_1(x)y(x) + a_0(x) \\ y(a) = y_a \ \text{ et } \ y(b) = y_b \end{array} \tag{7.29}$$

On suppose les fonctions $a_i(x)$ suffisamment régnlières pour assurer l'existence et l'unicité des équations différentielles que nous rencontrerons. La différence entre les équations différentielles avec conditions initiales et celles avec conditions aux limites est illustrée à la figure 7.5. Dans le premier cas, à $t = t_0$, la fonction $y(t_0)$ ainsi que sa pente $y'(t_0)$ sont connues. Dans le

cas des équations avec conditions aux limites, on ne connaît que les valeurs de la fonction $y(x)$ aux deux extrémités de l'intervalle, soit $y(a)$ et $y(b)$. On remarque qu'il n'y a aucune condition initiale liée à la dérivée de la fonction $y(x)$ en $x = a$. La condition initiale sur la dérivée est remplacée par une condition sur la fonction $y(x)$ à l'autre extrémité de l'intervalle $[a, b]$. On note également que la variable indépendante est maintenant notée x, car, contrairement au cas des équations différentielles avec conditions initiales, la variable indépendante des équations différentielles avec conditions aux limites est le plus souvent une variable d'espace, rarement de temps.

Cette incursion dans le domaine des équations différentielles est facilitée par le fait que nous pouvons résoudre l'équation 7.29 avec les outils que nous possédons déjà. Il suffit en effet de remarquer que l'on peut remplacer l'équation différentielle 7.29 par deux équations différentielles avec conditions initiales (voir Burden et Faires, réf. [2]).

Théorème 7.2

La solution de l'équation différentielle avec conditions aux limites 7.29 est donnée par:

$$y(x) = y_1(x) + \left(\frac{y_b - y_1(b)}{y_2(b)}\right) y_2(x) \tag{7.30}$$

où $y_1(x)$ et $y_2(x)$ sont les solutions des équations différentielles avec conditions initiales suivantes:

$$\begin{array}{c} y_1''(x) = a_2(x)y_1'(x) + a_1(x)y_1(x) + a_0(x) \\ y_1(a) = y_a \text{ et } y_1'(a) = 0 \end{array} \tag{7.31}$$

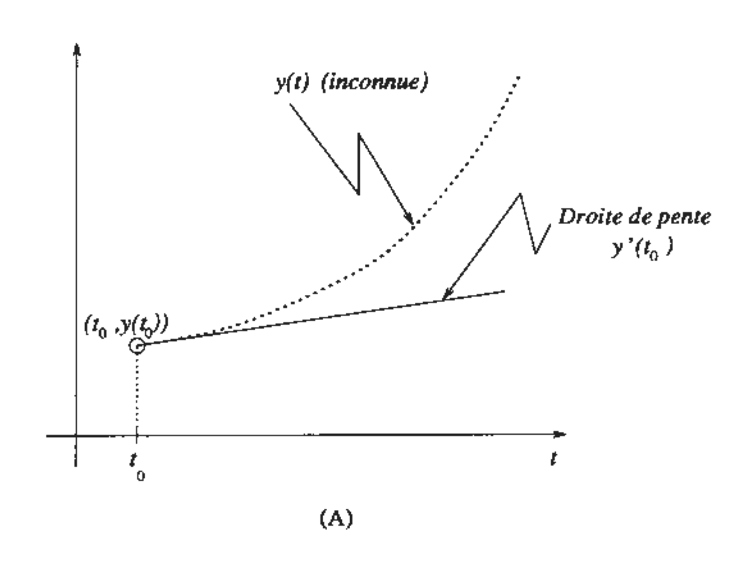
et

$$\begin{array}{c} y_2''(x) = a_2(x)y_2'(x) + a_1(x)y_2(x) \\ y_2(a) = 0 \text{ et } y_2'(a) = 1 \end{array} \tag{7.32}$$

Démonstration:

On doit vérifier en premier lieu les conditions aux limites. Si les fonctions $y_1(x)$ et $y_2(x)$ satisfont respectivement aux équations 7.31 et 7.32, en $x = a$ on a:

$$y(a) = y_1(a) + \left(\frac{y_b - y_1(b)}{y_2(b)}\right) y_2(a) = y_a + \left(\frac{y_b - y_1(b)}{y_2(b)}\right) 0 = y_a$$



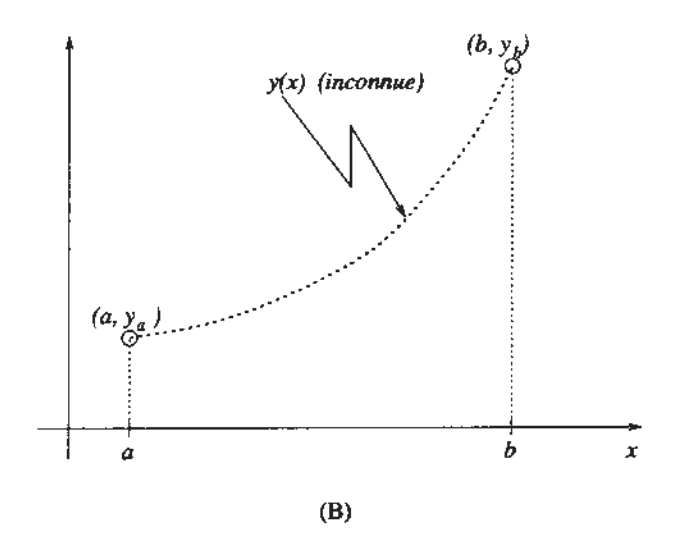


Figure 7.5: Conditions initiales (A) et aux limites (B)

Par ailleurs:

$$y(b) = y_1(b) + \left(\frac{y_b - y_1(b)}{y_2(b)}\right) y_2(b) = y_1(b) + (y_b - y_1(b)) = y_b$$

Il reste à montrer que l'expression 7.30 est bien la solution de l'équation 7.29. On pose:

$$c = \left(\frac{y_b - y_1(b)}{y_2(b)}\right)$$

pour simplifier la notation. On doit donc s'assurer que $y_1(x) + cy_2(x)$ est la solution de l'équation différentielle 7.29. La dérivée seconde de $y(x)$ peut alors s'écrire:

$$y''(x) = (y_1(x) + cy_2(x))'' = y_1''(x) + cy_2''(x)$$

Les fonctions $y_1(x)$ et $y_2(x)$ sont respectivement les solutions des équations 7.31 et 7.32. On a alors:

$$\begin{aligned}
y''(x) &= (a_2(x)y_1'(x) + a_1(x)y_1(x) + a_0(x)) \\
&\quad + c(a_2(x)y_2'(x) + a_1(x)y_2(x)) \\
&= a_2(x)(y_1'(x) + cy_2'(x)) + a_1(x)(y_1(x) + cy_2(x)) + a_0(x) \\
&= a_2(x)y'(x) + a_1(x)y(x) + a_0(x)
\end{aligned}$$

ce qui montre bien que $y(x)$, définie par 7.30, est la solution de l'équation 7.29. □

Remarque 7.19

On note l'absence du terme $a_0(x)$ de l'équation différentielle relative à $y_2(x)$. □

La figure 7.6 illustre les fonctions $y_1(x)$ et $y_2(x)$ ainsi que la solution recherchée. Si on regarde les conditions initiales relatives à $y_1(x)$ et à $y_2(x)$, on constate entre autres choses que ni $y_1(x)$ ni $y_2(x)$ ne vérifie la bonne condition aux limites en $x = b$. Par contre, en utilisant l'équation 7.30, on obtient la bonne solution.

On peut résoudre les équations différentielles avec conditions initiales 7.31 et 7.32 à l'aide de la méthode de Runge-Kutta d'ordre 4. On doit d'abord

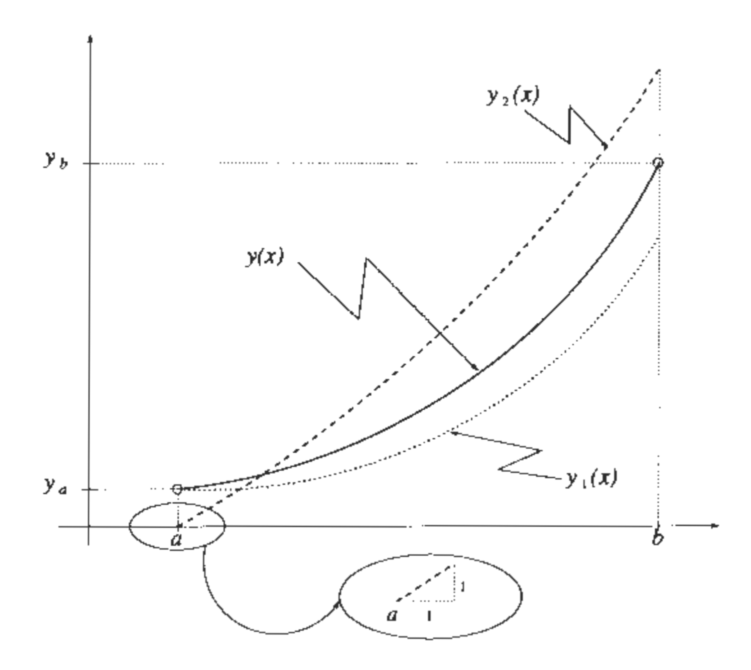


Figure 7.6: Méthode de tir

transformer chacune d'elles en un système de 2 équations différentielles d'ordre 1. En posant:

$$\begin{array}{lcl} u_1(x) = y_1(x) & \text{et} & v_1(x) = y_2(x) \\ u_2(x) = y_1'(x) & \text{et} & v_2(x) = y_2'(x) \end{array}$$

on obtient les deux systèmes suivants:

$$\begin{array}{lcll} u_1'(x) & = & u_2(x) & (u_1(a) = y_a) \\ u_2'(x) & = & a_2(x)u_2(x) + a_1(x)u_1(x) + a_0(x) & (u_2(a) = 0) \\ & & & \\ v_1'(x) & = & v_2(x) & (v_1(a) = 0) \\ v_2'(x) & = & a_2(x)v_2(x) + a_1(x)v_1(x) & (v_2(a) = 1) \end{array} \tag{7.33}$$

La solution finale peut alors s'écrire:

$$y(x) = y_1(x) + \left(\frac{y_b - y_1(b)}{y_2(b)}\right) y_2(x) = u_1(x) + \left(\frac{y_b - u_1(b)}{v_1(b)}\right) v_1(x)$$

selon les nouvelles variables $u_1(x)$ et $v_1(x)$.

Exemple 7.15
Soit l'équation différentielle suivante:

$$y''(x) = -\frac{2}{x}y'(x) + \frac{1}{x^2}$$

avec les conditions aux limites $y(1) = 0$ et $y(2) = 0{,}693\,147$. On a dans ce cas:

$$a_2(x) = -\frac{2}{x}, \quad a_1(x) = 0 \text{ et } a_0(x) = \frac{1}{x^2}$$

Le tableau qui suit présente la solution de cette équation différentielle. On peut démontrer facilement que la solntion analytique de cette équation est $y(x) = \ln x$, ce qui permet de calculer l'erreur absolue qui figure dans la dernière colonne du tableau.

x	$y_1(x) = u_1(x)$	$y_2(x) = v_1(x)$	$y(x)$	Erreur
1,0	0,000 000 00	0,000 000 00	0,000 000 00	$0{,}0000 \times 10^{-0}$
1,1	0,004 402 51	0,090 907 02	0,095 309 75	$0{,}4299 \times 10^{-6}$
1,2	0,015 656 98	0,166 663 59	0,182 320 96	$0{,}5997 \times 10^{-6}$
1,3	0,031 597 42	0,230 765 68	0,262 363 63	$0{,}6325 \times 10^{-6}$
1,4	0,050 760 44	0,285 710 54	0,336 471 64	$0{,}5924 \times 10^{-6}$
1,5	0,072 134 27	0,333 329 55	0,405 464 59	$0{,}5143 \times 10^{-6}$
1,6	0,095 006 09	0,374 996 26	0,470 003 21	$0{,}4173 \times 10^{-6}$
1,7	0,118 865 94	0,411 761 06	0,530 627 94	$0{,}3123 \times 10^{-6}$
1,8	0,143 344 54	0,444 440 91	0,587 786 46	$0{,}2057 \times 10^{-6}$
1,9	0,168 171 91	0,473 680 79	0,641 853 79	$0{,}1008 \times 10^{-6}$
2,0	0,193 149 33	0,499 996 70	0,693 147 18	$0{,}4400 \times 10^{-9}$

On a employé la méthode de Runge-Kutta d'ordre 4 pour le calcul de $y_1(x)$ et de $y_2(x)$. On note que:

$$y_1(b) = y_1(2{,}0) = 0{,}193\,149\,33 \text{ et } y_2(b) = y_2(2{,}0) = 0{,}499\,996\,70$$

ce qui permet le calcul de $y(x)$ à l'aide de l'équation 7.30.

● ● ● ● ●

On rencontre fréquemment un autre type d'équations différentielles avec conditions aux limites de la forme:

$$\begin{array}{c} y''(x) = a_2(x)y'(x) + a_1(x)y(x) + a_0(x) \\ y(a) = y_a \;\text{ et }\; y'(b) = y'_b \end{array} \tag{7.34}$$

La deuxième condition aux limites (c'est-à-dire en $x = b$) porte sur la dérivée. Un raisonnement similaire au précédent conduit au théorème suivant dont la démonstration est laissée en exercice.

Théorème 7.3

La solution de l'équation différentielle avec conditions aux limites 7.34 est donnée par:

$$y(x) = y_1(x) + \left(\frac{y'_b - y'_1(b)}{y'_2(b)}\right) y_2(x) \tag{7.35}$$

où $y_1(x)$ et $y_2(x)$ sont les solutions des équations différentielles avec conditions initiales 7.31 et 7.32. □

Remarque 7.20

Suivant la notation du système 7.33, la solution 7.35 s'écrit également:

$$y(x) = y_1(x) + \left(\frac{y'_b - y'_1(b)}{y'_2(b)}\right) y_2(x) = u_1(x) + \left(\frac{y'_b - u_2(b)}{v_2(b)}\right) v_1(x)$$

en fonction des variables des deux systèmes linéaires d'ordre 1. □

Exemple 7.16

Une tige métallique (voir Reddy, réf. [24]) de diamètre D = 0,02 m et de longueur L = 0,05 m est exposée à l'air ambiant à une température T_∞ = 20 °C. La première extrémité de la tige est maintenue à T_0 = 320 °C, tandis que l'autre extrémité est supposée parfaitement isolée, c'est-à-dire qu'il n'y a aucun flux de chaleur à cette extrémité ou encore que:

$$T'(L) = 0$$

La figure 7.7 résume la situation. L'équation différentielle régissant la tem-

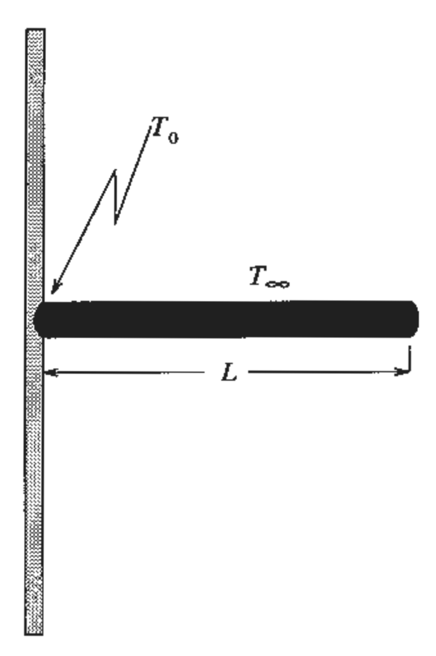


Figure 7.7: Transfert thermique dans une tige métallique

pérature $T(x)$ dans la tige est de la forme:

$$\theta''(x) = N^2\theta(x)$$

où $\theta(x) = T(x) - T_\infty$ et:

$$N^2 = \frac{4\beta}{kD}$$

Dans la définition de N, β est un coefficient de transfert de chaleur avec le milieu ambiant ($\beta = 100$ W$\cdot$ °C$^{-1}\cdot$ m^{-2}) et $k = 50$ W $\cdot$ °C$^{-1}\cdot$ m^{-1} est la conductivité thermique.
Les conditions aux limites relatives à θ sont donc:

$$\theta(0) = T(0) - 20\,^{\circ}\mathrm{C} = 300\,^{\circ}\mathrm{C} \;\text{ et }\; \theta'(0{,}05) = 0$$

ce qui correspond aux conditions du théorème précédent. Les valeurs des

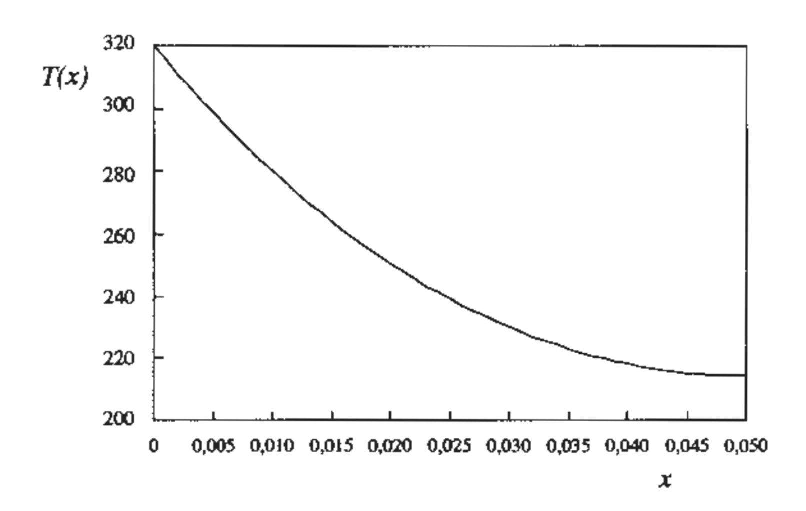


Figure 7.8: Température dans une tige métallique: $T(x) = \theta(x) + T_\infty$

différentes variables pour $h = 0{,}005$ sont les suivantes.

x	$y_1(x) = u_1(x)$	$y_2(x) = v_1(x)$	$\theta(x)$	Erreur
0,000	300,000 00	0,000 000 000	300,000 00	$0{,}0000 \times 10^{+0}$
0,005	301,501 25	0,005 008 333	278,615 36	$0{,}2918 \times 10^{-4}$
0,010	306,020 02	0,010 066 792	260,019 18	$0{,}5371 \times 10^{-4}$
0,015	313,601 54	0,015 226 002	244,025 37	$0{,}7417 \times 10^{-4}$
0,020	324,321 68	0,020 537 599	230,473 83	$0{,}9109 \times 10^{-4}$
0,025	338,287 74	0,026 054 743	219,228 95	$0{,}1048 \times 10^{-3}$
0,030	355,639 48	0,031 832 651	210,178 19	$0{,}1159 \times 10^{-3}$
0,035	376,550 59	0,037 929 150	203,230 95	$0{,}1244 \times 10^{-3}$
0,040	401,230 33	0,044 405 257	198,317 72	$0{,}1307 \times 10^{-3}$
0,045	429,925 71	0,051 325 786	195,389 31	$0{,}1350 \times 10^{-3}$
0,050	462,923 93	0,058 759 999	194,416 42	$0{,}1373 \times 10^{-3}$

Bien qu'elles ne figurent pas dans ce tableau, les valeurs:

$$y_1'(0{,}05) = 7051{,}1999 \quad \text{et} \quad y_2'(0{,}05) = 1{,}543\,079$$

obtenues lors de la résolution pour $y_1(x)$ et $y_2(x)$ sont nécessaires pour déterminer $y(x)$ en vertu de l'équation 7.35. Ici encore, une solution analytique existe:

$$\theta(x) = 300\frac{\cosh(N(L-x))}{\cosh(NL)}$$

qui permet d'évaluer l'erreur commise. La figure 7.8 présente la solution calculée. On remarque que les deux conditions aux limites sont bien vérifiées et en particulier que la pente de la température est nulle en $x = 0{,}05$, ce qui confirme que le flux de chaleur est nul.

●●●●●

7.9 Méthodes des différences finies

Dans cette section, nous nous intéressons uniquement aux équations différentielles avec conditions aux limites et nous proposons une solution de remplacement à la méthode de tir de la section précédente. Il s'agit de la *méthode des différences finies*, qui constitue la principale application des techniques de différentiation numérique que nous avons vues. De plus, cette méthode s'étend facilement aux équations aux dérivées partielles, ce qui n'est pas le cas de la méthode de tir.

On considère l'équation différentielle avec conditions aux limites:

$$\begin{gathered} y''(x) = a_2(x)y'(x) + a_1(x)y(x) + a_0(x) \\ y(a) = y_a \text{ et } y(b) = y_b \end{gathered}$$

que l'on réécrit sous la forme:

$$\begin{gathered} y''(x) - a_2(x)y'(x) - a_1(x)y(x) = a_0(x) \\ y(a) = y_a \text{ et } y(b) = y_b \end{gathered}$$

L'objectif est toujours de trouver une approximation y_i de $y(x_i)$ en certains points x_i de l'intervalle $[a, b]$. On divise d'abord cet intervalle en n sous-intervalles de longueur $h = (b - a)/n$ et on note $x_0 = a$, $x_i = a + ih$ et enfin $x_n = a + nh = b$. Les conditions aux limites imposent immédiatement que $y_0 = y_a$ et que $y_n = y_b$. Il reste donc à déterminer les $(n - 1)$ inconnues $y_1, y_2, \cdots, y_{n-1}$. Pour ce faire, on construit un système linéaire de dimension $(n - 1)$. La stratégie de résolution consiste à remplacer dans l'équation différentielle toutes les dérivées de la fonction $y(x)$ par des formules aux différences finies, et ce en chaque point x_i pour i allant de 1 à $(n - 1)$. Plus précisément, au point x_i, l'équation différentielle s'écrit:

$$y''(x_i) - a_2(x_i)y'(x_i) - a_1(x_i)y(x_i) = a_0(x_i)$$

On introduit ici les différences centrées:

$$y''(x_i) = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + O(h^2) \text{ et } y'(x_i) = \frac{y_{i+1} - y_{i-1}}{2h} + O(h^2)$$

pour obtenir:

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - a_2(x_i)\left(\frac{y_{i+1} - y_{i-1}}{2h}\right) - a_1(x_i)y_i = a_0(x_i) + O(h^2)$$

En négligeant le terme d'erreur $O(h^2)$ et en multipliant par $-2h^2$, on trouve:

$$-(2 + ha_2(x_i))\,y_{i-1} + (4 + 2h^2a_1(x_i))y_i + (-2 + ha_2(x_i))\,y_{i+1} = -2h^2a_0(x_i)$$

Cette dernière relation est vérifiée pour $i = 1, 2, \cdots, (n-1)$. On note de plus que, dans la première équation ($i = 1$), $y_{i-1} = y_0 = y_a$ et est une quantité connue. De même, à l'autre extrémité ($i = n-1$), on a $y_{i+1} = y_n = y_b$. On est ainsi amené à résoudre le système linéaire:

$$\begin{bmatrix} 4 + 2h^2a_1(x_1) & -2 + ha_2(x_1) & 0 & 0 \\ -2 - ha_2(x_2) & 4 + 2h^2a_1(x_2) & -2 + ha_2(x_2) & \vdots \\ 0 & \ddots & \ddots & \vdots \\ 0 & -2 - ha_2(x_{n-2}) & 4 + 2h^2a_1(x_{n-2}) & -2 + ha_2(x_{n-2}) \\ 0 & \cdots & -2 - ha_2(x_{n-1}) & 4 + 2h^2a_1(x_{n-1}) \end{bmatrix}$$

$$\times \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-2} \\ y_{n-1} \end{bmatrix} = \begin{bmatrix} -2h^2a_0(x_1) + (2 + ha_2(x_1))y_a \\ -2h^2a_0(x_2) \\ \vdots \\ -2h^2a_0(x_{n-2}) \\ -2h^2a_0(x_{n-1}) + (2 - ha_2(x_{n-1}))y_b \end{bmatrix} \tag{7.36}$$

Cette matrice est tridiagonale. Sa diagonale principale vaut $4 + 2h^2a_1(x_i)$ pour $i = 1, 2, \cdots, (n-1)$, sa diagonale inférieure vaut $-2 - ha_2(x_i)$ pour $i = 2, 3, \cdots, (n-1)$ et enfin sa diagonale supérieure vaut $-2 + ha_2(x_i)$ pour $i = 1, 2, \cdots, (n-2)$. Tous les autres termes sont nuls. La résolution numérique d'une telle matrice est très rapide. On remarque de plus que les première et dernière lignes tiennent compte des conditions aux limites, par le biais du terme de droite.

Exemple 7.17

On considère l'équation différentielle avec conditions aux limites:

$$y''(x) = -\frac{2}{x}y'(x) + \frac{1}{x^2}$$

$$y(0) = 0 \text{ et } y(1) = 0{,}693\,147$$

déjà résolue à l'aide de la méthode de tir. Dans ce cas, $a_2(x) = -2/x$, $a_1(x) = 0$ et $a_0(x) = 1/x^2$. Les conditions aux limites sont $y_a = 0$ et $y_b = 0{,}693\,147$. Si on prend 10 intervalles ($h = 0{,}1$), le système 7.36 est de dimension 9 et s'écrit (les coefficients de la matrice ont été tronqués à 3 chiffres):

$$\begin{bmatrix} 4{,}0 & -2{,}18 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ -1{,}83 & 4{,}00 & -2{,}16 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ 0{,}0 & -1{,}84 & 4{,}0 & -2{,}15 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ 0{,}0 & 0{,}0 & -1{,}85 & 4{,}0 & -2{,}14 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ 0{,}0 & 0{,}0 & 0{,}0 & -1{,}86 & 4{,}0 & -2{,}13 & 0{,}0 & 0{,}0 & 0{,}0 \\ 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & -1{,}87 & 4{,}0 & -2{,}12 & 0{,}0 & 0{,}0 \\ 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & -1{,}88 & 4{,}0 & -2{,}11 & 0{,}0 \\ 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & -1{,}88 & 4{,}0 & -2{,}11 \\ 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & -1{,}89 & 4{,}0 \end{bmatrix}$$

$$\times \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} -0{,}016\,53 \\ -0{,}013\,89 \\ -0{,}011\,83 \\ -0{,}010\,20 \\ -0{,}008\,89 \\ -0{,}007\,81 \\ -0{,}006\,92 \\ -0{,}006\,17 \\ +1{,}453\,72 \end{bmatrix}$$

On obtient les résultats suivants:

x_i	$y(x_i)$	y_i	$\lvert y(x_i) - y_i\rvert$
1,0	0,000 0000	0,000 0000	$0{,}000\,00 \times 10^{-0}$
1,1	0,095 3101	0,095 3410	$0{,}308\,38 \times 10^{-4}$
1,2	0,182 3215	0,182 3676	$0{,}460\,66 \times 10^{-4}$
1,3	0,262 3642	0,262 4157	$0{,}515\,12 \times 10^{-4}$
1,4	0,336 4722	0,336 5229	$0{,}507\,48 \times 10^{-4}$
1,5	0,405 4651	0,405 5111	$0{,}460\,29 \times 10^{-4}$
1,6	0,470 0036	0,470 0424	$0{,}388\,08 \times 10^{-4}$
1,7	0,530 6282	0,530 6582	$0{,}300\,39 \times 10^{-4}$
1,8	0,587 7866	0,587 8070	$0{,}203\,57 \times 10^{-4}$
1,9	0,641 8538	0,641 8640	$0{,}101\,88 \times 10^{-4}$
2,0	0,693 1470	0,693 1470	$0{,}000\,00 \times 10^{-0}$

On remarque immédiatement que l'erreur commise ici est plus grande que l'erreur produite par la méthode de tir. Cette différence d'erreur provient de

l'ordre de précision des méthodes employées. En effet, la méthode de tir est d'ordre 4 puisqu'elle s'appuie sur une méthode de Runge-Kutta d'ordre 4. Dans le cas de la méthode de différences finies, l'erreur commise est d'ordre 2 puisqu'on a utilisé des différences centrées d'ordre 2. On pourrait améliorer la précision en insérant par exemple des différences centrées d'ordre 4.

● ● ● ● ●

Remarque 7.21
Le choix de différences centrées, plutôt que de différences avant ou arrière, permet d'obtenir des résultats davantage précis, car d'ordre plus élevé. Toutefois, il peut s'avérer utile daus certaines situations d'utiliser des différences avant ou arrière lorsqu'on recherche des schémas aux différences finies quelque peu différents. □

7.10 Applications

7.10.1 Problème du pendule

Comme nous l'avons vu au chapitre 1, si un pendule de masse m et de longueur l fait un angle $\theta(t)$ avec la verticale, $\theta(t)$ est la solution de l'équation différentielle:

$$\theta''(t) = -\frac{c_f\theta'(t)}{m} - \frac{g\sin(\theta(t))}{l} \tag{7.37}$$

$$\theta(0) = \theta_0 \tag{7.38}$$

$$\theta'(0) = \theta'_0 \tag{7.39}$$

Le paramètre c_f est un coefficient de frottement qui est souvent négligé et $g = 9{,}8\ \mathrm{m/s^2}$. Il s'agit bien sûr d'une équation différentielle d'ordre 2 avec des conditions relatives à la position et à la vitesse angulaire initiales. Suivant les techniques de la section 7.7, on transforme cette équation d'ordre 2 en un système de 2 équations d'ordre 1. Ce système s'écrit, dans le cas d'un pendule *de masse et de longueur unitaires*:

$$\begin{array}{rcll} y'_1(t) & = & y_2(t) & (y_1(0) = \theta_0) \\ y'_2(t) & = & -c_f y_2(t) - g\sin(y_1(t)) & (y_2(0) = \theta'_0) \end{array} \tag{7.40}$$

Dans cet exemple, la variable $y_1(t)$ est l'angle $\theta(t)$ et la variable $y_2(t)$ est la vitesse angulaire $\theta'(t)$. On résout ce système à l'aide de la méthode de

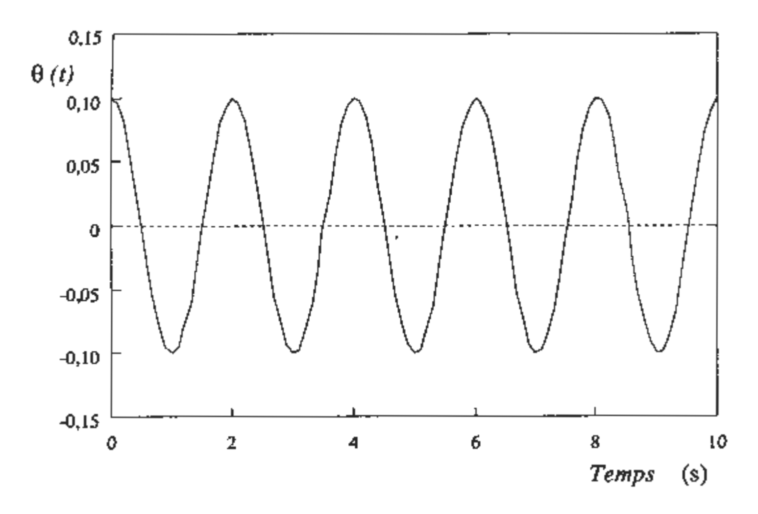


Figure 7.9: Pendule: $c_f = 0$, $\sin(\theta(t)) \simeq \theta(t)$ avec $\theta_0 = 0{,}1$ et $\theta_0' = 0$

Runge-Kutta d'ordre 4 et de l'algorithme 7.21. Dans un premier temps, on néglige le terme de frottement en posant $c_f = 0$, qui est une approximation fréqnemment utilisée. De plus, on fait appel à l'approximation $\sin(\theta(t)) \simeq \theta(t)$, qui est valide dans le cas où l'angle $\theta(t)$ reste petit en tout temps. Le système 7.40 devient alors:

$$\begin{array}{rcll} y_1'(t) & = & y_2(t) & (y_1(0) = \theta_0) \\ y_2'(t) & = & -gy_1(t) & (y_2(0) = \theta_0') \end{array}$$

Pour les conditions initiales $\theta_0 = 0{,}1$ rad et $\theta_0' = 0$, qui indiquent que la position initiale du pendule fait un angle de 0,1 rad avec la verticale et que sa vitesse angulaire initiale est nulle, on obtient les résultats de la figure 7.9.

Les conditions initiales choisies résultent en des oscillations de faible amplitude. Si on retire l'approximation $\sin(\theta(t)) \simeq \theta(t)$ tout en maintenant les mêmes conditions initiales, on doit résoudre:

$$\begin{array}{rcll} y_1'(t) & = & y_2(t) & (y_1(0) = \theta_0) \\ y_2'(t) & = & -g\sin(y_1(t)) & (y_2(0) = \theta_0') \end{array}$$

La figure 7.10 illustre la solution de ce nouveau système. Les courbes des figures 7.9 et 7.10 sont parfaitement superposables, ce qui démontre la validité de l'approximation $\sin(\theta(t)) \simeq \theta(t)$ pour ces conditions initiales.

Cette approximation n'est cependant plus valable dans le cas où les angles $\theta(t)$ deviennent grands. Ce cas est illustré à la figure 7.11, où on a retenu les

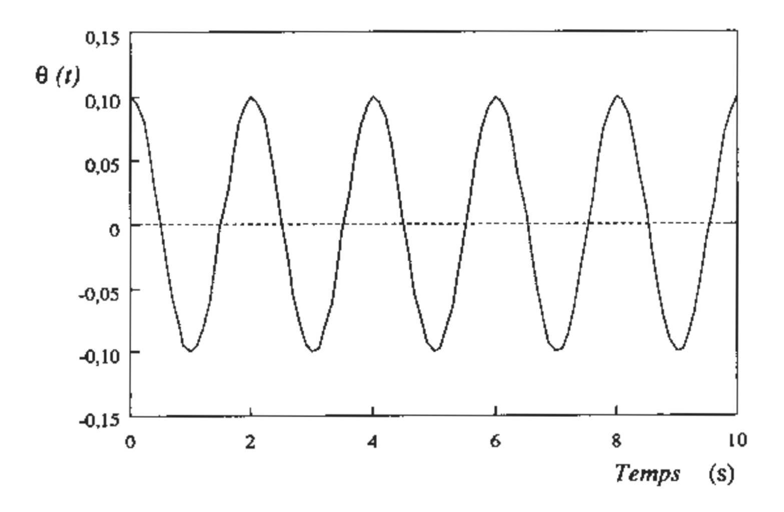


Figure 7.10: Pendule: $c_f = 0$ avec $\theta_0 = 0{,}1$ et $\theta_0' = 0$

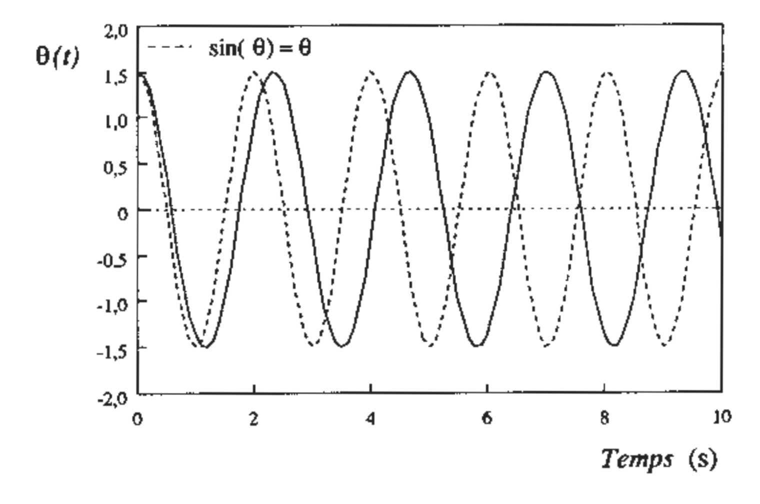


Figure 7.11: Pendule: $c_f = 0$ avec $\theta_0 = 1{,}5$ et $\theta_0' = 0$

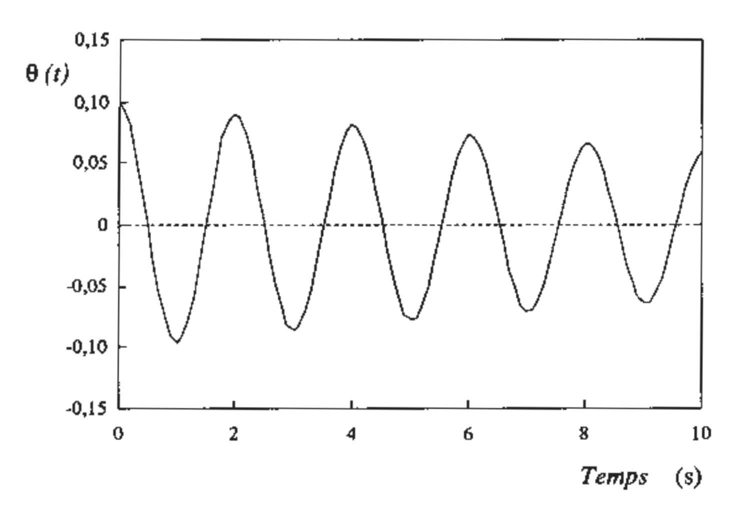


Figure 7.12: Pendule: $c_f = 0{,}1$ avec $\theta_0 = 0{,}1$ et $\theta_0' = 0$

conditions initiales $\theta_0 = 1{,}5$ rad et $\theta_0' = 0$ qui permettent des oscillations de forte amplitude. On voit nettement à la figure 7.11 que, bien que les deux courbes demeurent périodiques, elles diffèrent de manière importante.

Un dernier exemple permet de mettre en évidence l'influence de la force de frottement et du coefficient c_f. La figure 7.12 illustre le cas où $c_f = 0{,}1$. On voit très nettement l'amplitude des oscillations diminuer. Le pendule finit par s'arrêter complètement. Un coefficient c_f plus grand amortit plus rapidement les oscillations.

7.10.2 Systèmes de masses et de ressorts

Prenons comme seconde application un système de deux masses et de trois ressorts, tel que l'illustre la figure 7.13 (voir Strang, réf. [21]). La position d'équilibre est atteinte lorsque les forces de tension dans les ressorts sont équilibrées par le poids des masses.

On note m_i, les masses et c_i, les coefficients de rigidité des ressorts. La *loi de Hooke* établit une relation linéaire entre la force f_i et l'élongation e_i du i^e ressort de la forme:

$$f_i = c_i e_i \tag{7.41}$$

Si on perturbe l'équilibre en déplaçant une ou plusieurs masses, le système se met à vibrer suivant les équations:

$$M\frac{d^2\vec{u}}{dt^2} = -K\vec{u}$$

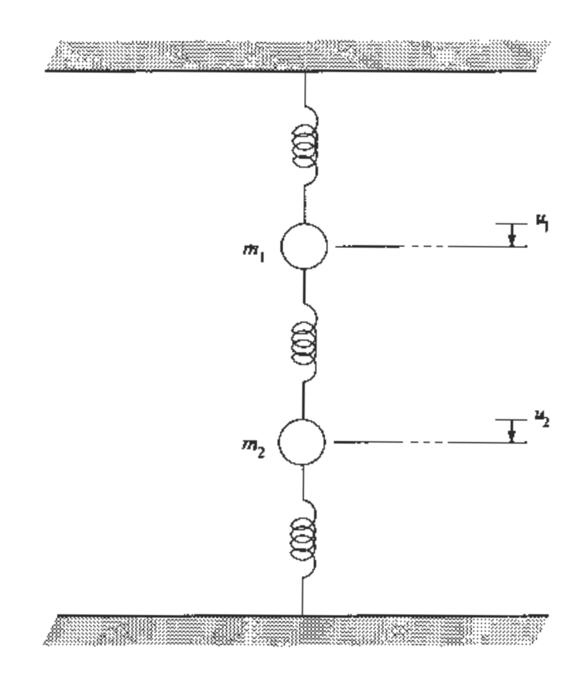


Figure 7.13: Système de masses et de ressorts

ou encore

$$\frac{d^2\vec{u}}{dt^2} = -M^{-1}K\vec{u} \tag{7.42}$$

où le vecteur $\vec{u}(t) = [u_1(t)\ u_2(t)]^T$ exprime le déplacement de chacune des masses par rapport à la position d'équilibre. Le vecteur vitesse est bien sûr noté $\vec{u}'(t)$. Les conditions initiales $\vec{u}(0) = [u_{1,0}\ u_{2,0}]^T$ et $\vec{u}'(0) = [u'_{1,0}\ u'_{2,0}]^T$ complètent le système. La matrice M est la *matrice de masse* définie par:

$$M = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix}$$

K est la *matrice de rigidité* définie par:

$$K = A^T C A$$

où A est la *matrice d'incidence*:

$$A = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}$$

et C est la matrice:

$$C = \begin{bmatrix} c_1 & 0 & 0 \\ 0 & c_2 & 0 \\ 0 & 0 & c_3 \end{bmatrix}$$

La matrice d'incidence A est construite de telle sorte que:

$$A\vec{u} = \vec{e}$$

En d'autres mots, pour des déplacements $\vec{u}$ donnés, $A\vec{u}$ donne l'élongation $\vec{e}$ de chaque ressort.

Le système 7.42 est en fait le résultat de l'application de la loi de Hooke 7.41 et du second principe de Newton. Dans le cas où les masses m_i et les constantes de rigidité c_i sont unitaires, les matrices M et C deviennent des matrices identité et la matrice K est simplement égale à A^TA, c'est-à-dire:

$$K = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

Le système 7.42 devient alors:

$$\begin{array}{rcll} u_1''(t) & = & -2u_1(t) + u_2(t) & (u_1(0) = u_{1,0} \text{ et } u_1'(0) = u_{1,0}') \\ u_2''(t) & = & u_1(t) - 2u_2(t) & (u_2(0) = u_{2,0} \text{ et } u_2'(0) = u_{2,0}') \end{array}$$

où $u_{i,0}$ et $u_{i,0}'$ sont respectivement la position et la vitesse initiales de la i^e masse. Il faut maintenant ramener ce système de 2 équations différentielles d'ordre 2 à un système de 4 équations différentielles d'ordre 1. Il suffit pour ce faire de poser:

$$\begin{array}{rcl} y_1(t) & = & u_1(t) \\ y_2(t) & = & u_1'(t) \\ y_3(t) & = & u_2(t) \\ y_4(t) & = & u_2'(t) \end{array}$$

et le système devient:

$$\begin{array}{rcll} y_1'(t) & = & y_2(t) & (y_1(0) = u_{1,0}) \\ y_2'(t) & = & -2y_1(t) + y_3(t) & (y_2(0) = u_{1,0}') \\ y_3'(t) & = & y_4(t) & (y_1(0) = u_{2,0}) \\ y_4'(t) & = & y_1(t) - 2y_3(t) & (y_2(0) = u_{2,0}') \end{array}$$

La figure 7.14 illustre les déplacements de chacune des deux masses en fonction du temps selon les conditions initiales $\vec{u}(0) = [0{,}1 \;\; 0{,}0]^T$ et $\vec{u}'(0) = [0{,}0 \;\; 0{,}0]^T$ qui indiquent qu'une seule des masses est initialement déplacée.

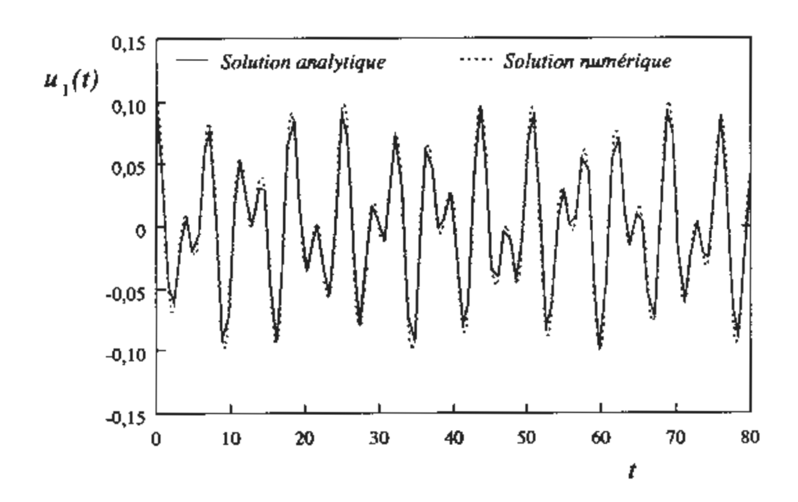


Figure 7.14: Déplacement $u_1(t)$ de la masse 1

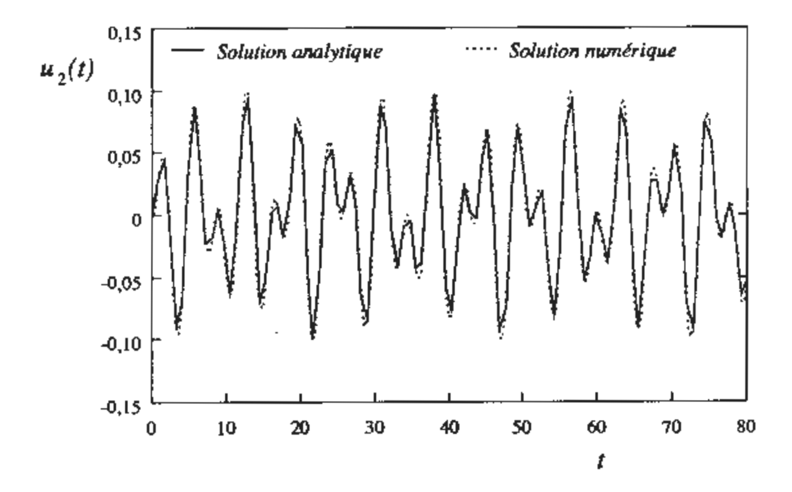


Figure 7.15: Déplacement $u_2(t)$ de la masse 2

On remarque le caractère non périodique de chacun des déplacements. On pourrait montrer que la solution analytique est de la forme:

$$\begin{aligned} u_1(t) = y_1(t) &= 0{,}05(\cos t + \cos\sqrt{3}t) \\ u_2(t) = y_3(t) &= 0{,}05(\cos t - \cos\sqrt{3}t) \end{aligned}$$

Cette solution est une combinaison de deux signaux dont les fréquences respectives sont 1 et $\sqrt{3}$. Ces fréquences sont dites *incommensurables*, ce qui signifie que leur rapport est un nombre irrationnel. C'est ce qui explique la non-périodicité de la solution.

7.10.3 Attracteur étrange de Lorenz

On connaît l'importance accordée aux prévisions météorologiques dans la vie de tous les jours. Il importe en effet de pouvoir prédire les conditions climatiques afin de planifier bon nombre d'activités. Ce besoin est encore plus aigu lorsque se préparent des ouragans ou de simples tempêtes de neige.

Or, la prévision météorologique est largement basée sur les méthodes numériques. Les travaux dans ce domaine remontent à quelques dizaines d'années. Un des pionniers fut Lorenz (réf. [17]), qui mit au point un modèle très simple de la forme:

$$\begin{aligned} y_1'(t) &= \sigma(y_2(t) - y_1(t)) \\ y_2'(t) &= r y_1(t) - y_2(t) - y_1(t) y_3(t) \\ y_3'(t) &= y_1(t) y_2(t) - b y_3(t) \end{aligned} \tag{7.43}$$

Selon Gulick (réf. [13]), la variable $y_1(t)$ est proportionnelle à l'intensité des mouvements de convection, la variable $y_2(t)$ est proportionnelle à la différence de température entre les courants ascendants et descendants, tandis que la variable $y_3(t)$ est proportionnelle à la distorsion du profil de température vertical par rapport au profil linéaire. Les paramètres sont $\sigma = 10$, $r = 2{,}666\,6667$ et $b = 28$.

Lors de l'introduction de ce modèle, Lorenz s'est livré malgré lui à une expérience numérique que nous allons tâcher de reproduire ici. À partir de la condition initiale $[y_1(0)\ y_2(0)\ y_3(0)]^T = [1\ 0\ 0]^T$ (on obtiendrait des résultats similaires avec toute autre condition initiale) et d'un pas de temps $h = 0{,}01$, Lorenz entreprit une simulation numérique qui devait produire une prévision météorologique sur plusieurs jours. Cependant, comme cela

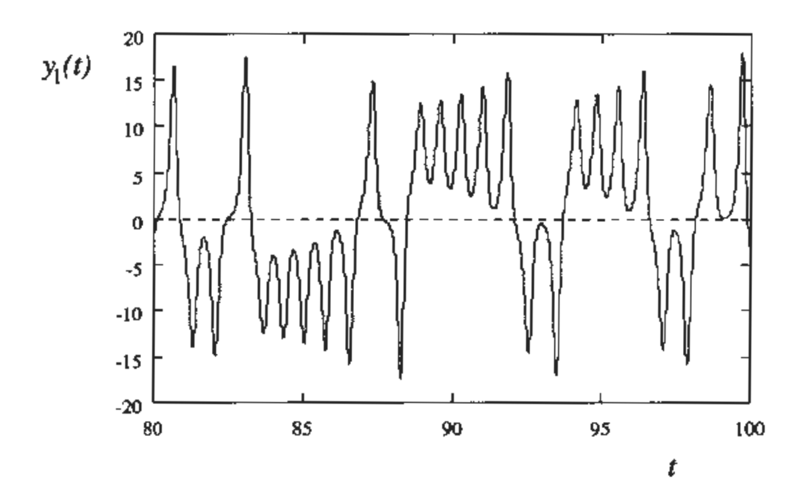


Figure 7.16: Lorenz: première simulation

arrive souvent dans ce type de calcul, une malheureuse panne d'électricité se produisit autour de $t = 100$. Plutôt que de reprendre les calculs à partir du début – les ordinateurs de l'époque étaient beaucoup plus lents que ceux d'aujourd'hui –, Lorenz choisit de reprendre la simulation à partir de $t = 80$ correspondant à la dernière solution imprimée par son programme. Cela lui permettait de comparer les deux simulations sur une période d'environ 2 000 pas de temps, c'est-à-dire entre $t = 80$ et $t = 100$, et de vérifier ainsi si tout se passait bien. La nouvelle condition initiale était $[y_1(80)\ y_2(80)\ y_3(80)]^T = [-2{,}4881\ \ 1{,}5045\ \ 26{,}865]^T$.

Lorenz constata à sa grande surprise que les deux solutions n'étaient pas du tout identiques. La figure 7.16 montre les résultats de la première simulation pour les 2 000 derniers pas de temps (les 8 000 premiers pas de temps ne sont pas illustrés). De son côté, la figure 7.17 montre les 2 000 premiers pas de temps à partir de la condition initiale $[-2{,}4881\ 1{,}5045\ 26{,}865]^T$. On constate aisément que les deux courbes sont similaires au début mais qu'elles diffèrent totalement par la suite. La figure 7.18 présente la superposition des deux courbes.

Il fallait donner une explication à ce curieux phénomène. Après plusieurs jours, on comprit que le format suivant lequel le programme écrivait la solution dans un fichier ne comptait que 5 chiffres significatifs, alors que les calculs étaient effectués avec l'équivalent de 7 ou 8 chiffres (en simple précision). La solution de la première simulation à $t = 80$ qu'aurait dû imprimer l'ordinateur était $[y_1(80)\ y_2(80)\ y_3(80)]^T = [-2{,}488\,1258\ 1{,}504\,5223\ 26{,}865\,757]^T$.

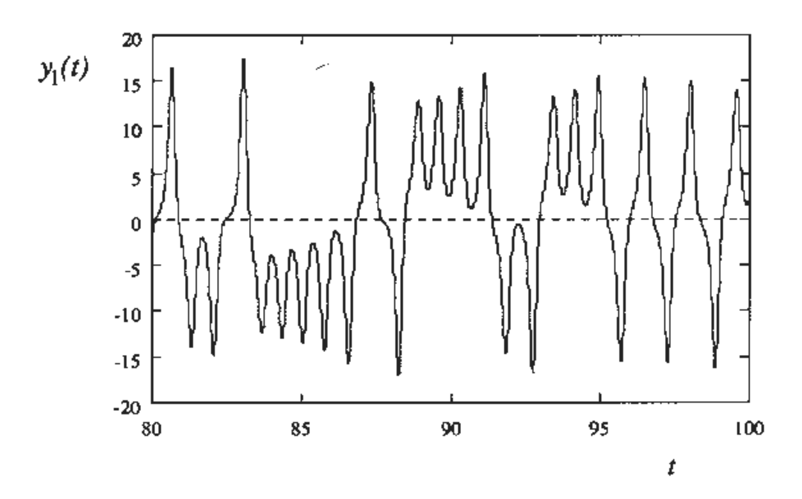


Figure 7.17: Lorenz: deuxième simulation

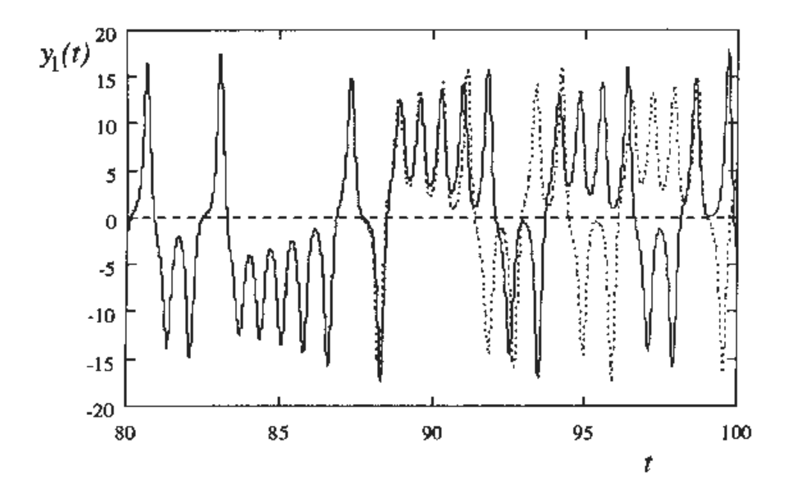


Figure 7.18: Lorenz: superposition des deux courbes

Il y avait donc une légère imprécision dans la condition initiale de la deuxième simulation. Ainsi, une petite erreur dans la connaissance de la condition initiale entraîne une très grande erreur dans la prévision numérique.

Une telle sensibilité par rapport anx conditions initiales est dramatique en ce qui concerne la prévision météorologique. En effet, les conditions initiales pour tous les modèles nnmériqnes de prévision sont obtenues à partir d'observations et de mesures dont la précision n'est pas parfaite. Il y a donc toujours une erreur dans les conditions initiales du modèle. L'expérience de Lorenz démontre que, dans ces conditions, la prévision météorologique à court terme est possible, mais qu'il en est tout autrement des prévisions à moyen et à long terme. En effet, après un temps relativement court, l'imprécision relative aux conditions initiales domine la simulation. C'est ce qu'on appelle l'*effet papillon*. Le simple battement des ailes d'un papillon à Tokyo pourrait être suffisant pour provoquer une tempête tropicale aux Antilles...

Terminons ce cas d'application en présentant le *plan de phase* que l'on obtient en traçant dans l'espace à 3 dimensions les points $(y_1(t), y_2(t), y_3(t))$. Il s'agit de l'*attracteur étrange de Lorenz*. Les plans de phase sont souvent utilisés pour visualiser l'interaction des différentes variables d'un système entre elles. Il s'agit de la trajectoire tracée par la solution $(y_1(t), y_2(t), y_3(t))$ à partir de la condition initiale. On obtiendrait un résultat semblable en partant de toute autre condition initiale. C'est pourquoi on qualifie cet attracteur d'étrange. Quelle que soit la condition initiale, la figure finale sera toujours celle de la figure 7.19. Les trajectoires sont donc invariablement attirées par cet ≪ objet ≫ que l'on nomme attracteur. Par ailleurs, si on considère deux conditions initiales très voisines, les deux trajectoires seront attirées par l'attracteur de Lorenz, mais elles suivront chacune un parcours totalement différent. C'est pourquoi on dit que l'attracteur est étrange. C'est cette extrême sensibilité aux conditions initiales qui crée l'effet papillon et qui rend ce type de problèmes tout à fait passionnant.

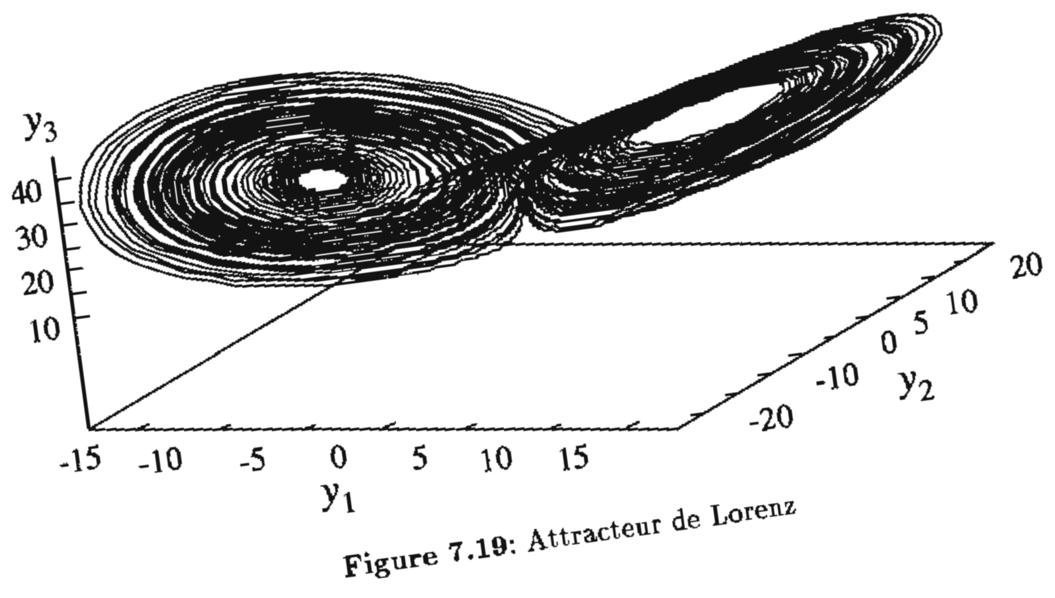


Figure 7.19: Attracteur de Lorenz

7.11 Exercices

1. Faire trois itérations avec $h = 0{,}1$ des méthodes d'Euler, d'Euler modifiée, du point milieu et de Runge-Kutta d'ordre 4 pour les équations différentielles suivantes:

$$\begin{array}{lll} \text{a)} & y'(t) = t\sin(y(t)) & (y(0) = 2) \\ \text{b)} & y'(t) = t^2 + (y(t))^2 + 1 & (y(1) = 0) \\ \text{c)} & y'(t) = y(t)e^t & (y(0) = 2) \end{array}$$

2. L'équation différentielle:

$$y'(t) = y(t) + e^{2t} \quad (y(0) = 2)$$

possède la solution analytique:

$$y(t) = e^t + e^{2t}$$

a) En prenant $h = 0{,}1$, faire 3 itérations de la méthode d'Euler modifiée et calculer l'erreur commise sur y_3 en comparant les résultats avec la solution analytique $y(0{,}3)$.

b) En prenant $h = 0{,}05$, faire 6 itérations de la méthode d'Euler modifiée et calculer l'erreur commise sur y_6 en comparant les résultats avec la solution analytique $y(0{,}3)$.

c) Faire le ratio des erreurs commises en a) et en b), et commenter le résultat en fonction de l'erreur de troncature locale liée à la méthode utilisée.

d) Utiliser l'extrapolation de Richardson pour obtenir une meilleure approximation de $y(0{,}3)$.

3. Refaire l'exercice précédent, mais cette fois à l'aide de la méthode de Runge-Kutta d'ordre 4.

4. Montrer que l'ordre de l'erreur de troncature locale de la méthode du point milieu est 2. Identifier en premier lieu la fonction $\phi(t, y(t))$.

5. Faire deux itérations de la méthode de Runge-Kutta d'ordre 4 (en prenant $h = 0{,}1$) pour le système d'équations différentielles suivant:

$$\begin{array}{ll} y_1'(t) = y_2(t) + y_1(t) & (y_1(1) = 2) \\ y_2'(t) = y_1(t) + t & (y_2(1) = 1) \end{array}$$

6. Transformer les équations différentielles d'ordre supérieur suivantes en systèmes d'équations différentielles d'ordre 1.

   a) $y^{(3)}(t) = y^{(2)}(t) + y^{(1)}(t) - y(t) + 1$
   $y(0) = 2,\ y^{(1)}(0) = 2 \text{ et } y^{(2)}(0) = 1$

   b) $y^{(2)}(t) = t^2 + (y(t))^2 + 1$
   $y(1) = 0 \text{ et } y^{(1)}(1) = 2$

   c) $y^{(4)}(t) = y^{(2)}(t)e^t + (y^{(3)}(t))^3$
   $y(0) = 2,\ y^{(1)}(0) = 1,\ y^{(2)}(0) = 0 \text{ et } y^{(3)}(0) = 4$

7. On doit résoudre l'équation différentielle d'ordre 2:

$$y^{(2)}(x) + y^{(1)}(x) - 2y(x) + 16 = 0$$

   avec $y(0) = -7$ et $y(3) = 12{,}0855$, à l'aide d'une méthode de tir.

   a) Donner l'expression des deux équations différentielles d'ordre 2 avec conditions initiales qui permettent de résoudre cette équation différentielle.

   b) Ramener chacune des équations différentielles avec conditions initiales obtenues en a) à un système de deux équations différentielles d'ordre 1.

   c) Décrire brièvement une stratégie de résolution numérique.

8. Montrer que l'équation 7.35 est bien une solution de l'équation différentielle 7.34.

9. Montrer que l'équation différentielle avec conditions aux limites:

$$\begin{gathered} y''(x) = a_2(x)y'(x) + a_1(x)y(x) + a_0(x) \\ y'(a) = y'_a \text{ et } y(b) = y_b \end{gathered}$$

   possède la solution:

$$y(x) = y_1(x) + \frac{y_b - y_1(b)}{y_2(b)} y_2(x)$$

   où $y_1(x)$ et $y_2(x)$ sont les solutions de:

$$\begin{gathered} y_1''(x) = a_2(x)y_1'(x) + a_1(x)y_1(x) + a_0(x) \\ y_1(a) = 0 \text{ et } y_1'(a) = y'_a \end{gathered}$$

et

$$y_2''(x) = a_2(x)y_2'(x) + a_1(x)y_2(x)$$
$$y_2(a) = 1 \text{ et } y_2'(a) = 0$$

10. Résoudre l'équation différentielle:

$$y'(t) = t\sin(y(t)) \quad (y(0) = 2)$$

à l'aide des méthodes de prédiction-correction d'ordre 2 et 4 (prendre $h = 0{,}1$ et utiliser les valeurs calculées à l'exercice 1 a) à l'aide de la méthode de Runge-Kutta d'ordre 4 pour obtenir les premières valeurs des y_i). Effectuer 3 itérations.

11. Donner le système tridiagonal requis pour résoudre l'équation différentielle:

$$y''(x) = \left(1 + \frac{2}{x}\right) y(x) - (x+2)$$
$$y(0) = 0 \text{ et } y(1) = 2$$

à l'aide de la méthode des différences finies centrées (prendre 5 intervalles).

12. On veut résoudre par la méthode des différences finies l'équation différentielle:

$$y''(x) - a_2(x)y'(x) - a_1(x)y(x) = a_0(x)$$

avec les conditions aux limites $y(a) = y_a$ et $y(b) = y_b$. Déterminer le système tridiagonal résultant lorsqu'on utilise une différence arrière d'ordre 1 pour $y'(x)$ et une différence centrée d'ordre 2 pour $y''(x)$. Quel est l'ordre de précision de cette méthode?

Réponses aux exercices

Réponses aux exercices du chapitre 1

1. a) $\Delta X \leq 0{,}5 \times 10^{-4}$, $E_r \simeq 0{,}405 \times 10^{-3}$

   b) $\Delta X \leq 0{,}5 \times 10^{-3}$, $E_r \simeq 0{,}570 \times 10^{-4}$

   c) $\Delta X \leq 0{,}5 \times 10^{-5}$, $E_r \simeq 0{,}159 \times 10^{-5}$

   d) $\Delta X \leq 0{,}5 \times 10^{-8}$, $E_r \simeq 0{,}445 \times 10^{-4}$

   e) $\Delta X \leq 0{,}5 \times 10^{-3}$, $E_r \simeq 0{,}625 \times 10^{-4}$

   f) $\Delta X \leq 0{,}5 \times 10^{+3}$, $E_r \simeq 0{,}223 \times 10^{-4}$

2. a) $(32)_{10} = (100\,000)_2$

   b) $(125)_{10} = (1\,111\,101)_2$

   c) $(1231)_{10} = (10\,011\,001\,111)_2$

   d) $(876)_{10} = (1\,101\,101\,100)_2$

   e) $(999)_{10} = (1\,111\,100\,111)_2$

   f) $(12\,345)_{10} = (11\,000\,000\,111\,001)_2$

3. a) Signe et grandeur: $(+125)_{10} = (01\,111\,101)_2$, $(-125)_{10} = (11\,111\,101)_2$, $(+0)_{10} = (00\,000\,000)_2$ ou $(-0)_{10} = (10\,000\,000)_2$, $(\pm 175)_{10}$ ne peuvent pas être représentés sur 8 bits, $(-100)_{10} = (11\,100\,100)_2$

   b) Complément à 2: $(+125)_{10} = (01\,111\,101)_2$, $(-125)_{10} = (10\,000\,011)_2$, $(0)_{10} = (00\,000\,000)_2$, $(\pm 175)_{10}$ ne peuvent pas être représentés sur 8 bits, $(-100)_{10} = (10\,011\,100)_2$

   c) Représentation par excès ($d = 2^7$): $(+125)_{10} = (11\,111\,101)_2$, $(-125)_{10} = (00\,000\,011)_2$, $(0)_{10} = (10\,000\,000)_2$, $(\pm 175)_{10}$ ne peuvent pas être représentés sur 8 bits, $(-100)_{10} = (00\,011\,100)_2$

4. a) Binaire classique: $(00\,000\,011)_2 = (3)_{10}$, $(10\,000\,001)_2 = (129)_{10}$, $(11\,111\,111)_2 = (255)_{10}$

   b) Signe et grandeur: $(00\,000\,011)_2 = (+3)_{10}$, $(10\,000\,001)_2 = (-1)_{10}$, $(11\,111\,111)_2 = (-127)_{10}$

   c) Complément à 2: $(00\,000\,011)_2 = (+3)_{10}$, $(10\,000\,001)_2 = (-127)_{10}$, $(11\,111\,111)_2 = (-1)_{10}$

   d) Représentation par excès ($d = 2^7$): $(00\,000\,011)_2 = (-125)_{10}$, $(10\,000\,001)_2 = (+1)_{10}$, $(11\,111\,111)_2 = (+127)_{10}$

5. a) Plus grand nombre: $(0\,011\,111\,111\,111\,111)_2 = (32\,736)_{10}$. Plus petit nombre: $(0\,100\,001\,000\,000\,000)_2 = (2^{-17})_{10}$

   b) $\epsilon = 2^{1-10} = 2^{-9}$

6. a) $(0{,}5)_{10} = (0{,}1)_2$

   b) $(0{,}2)_{10} = (0{,}001\,100\,110\,011\cdots)_2$

   c) $(0{,}9)_{10} = (0{,}111\,001\,100\,110\,011\cdots)_2$

   d) $(1/3)_{10} = (0{,}010\,101\,010\cdots)_2$

   e) $(0{,}25)_{10} = (0{,}01)_2$

   f) $(3/8)_{10} = (0{,}011)_2$

7. a) $-1345{,}0$

   b) 2^{1-24}

   c) $0111\,1111\,1000\,0000\,0000\,0000\,0000\,0000 = 2^{-64}$

8. a) 1100 0010 0101 0000 1111 0000 0000 0000

   b) 0100 0101 1101 1110 0100 0000 0000 0000

   c) 0100 0001 1000 0001 1001 1001 1001 1010 (valeur arrondie)

   En retransformant ces réponses en décimal, on trouve:

   a)$-52{,}234\,375$, b) $7\,112{,}0$ et c) $16{,}200\,000\,75$.

   Les deux premières représentations sont exactes, mais la troisième comporte une erreur absolue de $0{,}75 \times 10^{-6}$.

9. a) $e \to 0{,}2718 \times 10^1$

   b) $1/6 \to 0{,}1667 \times 10^0$

   c) $2/3 \to 0{,}6667 \times 10^0$

d) $12{,}487 \times 10^5 \rightarrow 0{,}1249 \times 10^7$

e) $213\,456 \rightarrow 0{,}2135 \times 10^6$

f) $2000{,}1 \rightarrow 0{,}2000 \times 10^4$

10. $(x + y) + z = 0{,}196 \times 10^4$, alors que $x + (y + z) = 0{,}195 \times 10^4$

11. a) $0{,}999 \times 10^0$

    b) $0{,}214 \times 10^8$

    c) $0{,}237 \times 10^1$

    d) $0{,}105 \times 10^5$

    e) $0{,}700 \times 10^3$

    f) $0{,}316 \times 10^4$

12. $9 \times 10 \times 10$ possibilités pour la mantisse, 19 exposants différents et 2 signes ($\pm$) pour un total de 34 200 nombres.

13. Non (voir l'exercice 9a) pour un exemple)

14. Si x est près de 1, $\cos x$ est également près de 1 et il y a risque d'élimination par soustraction des chiffres significatifs. Une solution de rechange est:

$$(1 - \cos x)\frac{(1 + \cos x)}{(1 + \cos x)} = \frac{\sin^2 x}{(1 + \cos x)}$$

15. a) $\cos 2\theta$

    b) Horner: $p(x) = 1 + x(-2 + x(3 - 4x))$

    c) On effectue la somme à rebours.

16. Si n est assez grand, la représentation en notation flottante de $1/n$ devient nulle et la série s'arrête.

17. $\frac{\Delta(XY)}{|XY|} = \frac{|X|\Delta Y + |Y|\Delta X}{|XY|} = \frac{\Delta X}{|X|} + \frac{\Delta Y}{|Y|}$

18. a) $\cos(0 + h) = 1 - \frac{h^2}{2!} + \frac{h^4}{4!} - \frac{h^6}{6!} + \frac{\cos(\xi(h))h^8}{8!}$ avec $0 \leq \xi(h) \leq h$

    b) $\sin(0 + h) = h - \frac{h^3}{3!} + \frac{h^5}{5!} - \frac{h^7}{7!} + \frac{\cos(\xi(h))h^9}{9!}$ avec $0 \leq \xi(h) \leq h$

    c) $\arctan(0 + h) = h - \frac{2h^3}{3!} + O(h^5)$ avec $0 \leq \xi(h) \leq h$

    d) $\cos(\pi/2 + h) = -h + \frac{h^3}{3!} - \frac{h^5}{5!} + \frac{\sin(\xi(h))h^7}{7!}$ avec $\pi/2 \leq \xi(h) \leq \pi/2 + h$

    e) $\sin(\pi/2 + h) = 1 - \frac{h^2}{2!} + \frac{h^4}{4!} - \frac{h^6}{6!} + \frac{\sin(\xi(h))h^8}{8!}$ avec $\pi/2 \leq \xi(h) \leq \pi/2 + h$

19. a) $\Delta f = 0{,}248 \times 10^{-2}$, $f(x^*) = 0{,}698\,13$ (2 chiffres significatifs)
    b) $\Delta f = 0{,}247 \times 10^{-4}$, $f(x^*) = 0{,}790\,37$ (4 chiffres significatifs)
    c) $\Delta f = 0{,}9 \times 10^{-2}$, $f(x^*) = 2{,}529\,519$ (2 chiffres significatifs)
    d) $\Delta f = 0{,}109 \times 10^{-2}$, $f(x^*) = 0{,}012\,0512$ (1 chiffre significatif)

20. a) $\Delta f = 0{,}6538 \times 10^{2}$, $f(x^*, y^*) = 7\,543{,}098$ (1 chiffre significatif)
    b) $\Delta f = 0{,}82 \times 10^{-3}$, $f(x^*, y^*) = -0{,}008\,1459$ (0 chiffre significatif)

21. a) Respectivement 5 (25,312) et 6 (25,3125) chiffres significatifs
    b) $e_1/e_2 = 16 = 2^n$, ordre 4

22. a) $\ln(1+h) = h - \frac{h^2}{2} + \frac{h^3}{3} - \frac{h^4}{4} + \frac{h^5}{5\xi^5}$
    b) $\ln(1{,}1) \simeq 0{,}095\,308\,333$ avec 4 chiffres significatifs
    c) On divise h par 4, ce qui revient à diviser l'erreur par 4^5.

23. C'est un développement d'ordre 7.

24. a) $f(x) = 1 + x + x^2 + x^3 + x^4 + \cdots$
    b) $g(t) = 1 - t^2 + t^4 - t^6 + t^8 - \cdots$
    c) $\arctan(t) = t - t^3/3 + t^5/5 - t^7/7 - \cdots$
    d) $\ln(1+x) = x - x^2/2 + x^3/3 - x^4/4 - \cdots$

25. a) $e^{-x} = 1 - x + x^2/2! - x^3/3! + x^4/4! - x^5/5! + \cdots$
    b) $e^{-t} = 1 - t^2 + t^4/2! - t^6/3! + t^8/4! - t^{10}/5! + \cdots$
    c) $f(x) = (2/\sqrt{\pi})(x - x^3/3 + x^5/10 - x^7/42 + x^9/216 + \cdots)$
    d) $f(1) \simeq (2/\sqrt{\pi}) = (1 - 1/3 + 1/10 - 1/42) = 0{,}838\,224\,524$
    e) C'est une approximation d'ordre 9.
    f) 2 chiffres significatifs

Réponses aux exercices du chapitre 2

1. a) $x_m = 2{,}9,\ 2{,}85,\ 2{,}875$ (9 itérations)
   b) $x_m = 1{,}75,\ 1{,}625,\ 1{,}6875$ (10 itérations)
   c) $x_m = 2{,}0,\ 3{,}0,\ 3{,}5$ (13 itérations)
   d) $x_m = 1{,}5,\ 1{,}25,\ 1{,}125$ (11 itérations)

2. $x_m^* = x_1 - \frac{f(x_1)(x_2-x_1)}{f(x_2)-f(x_1)}$

3. a) 2,857 95, 2,860 03, 2,860 10

   b) 1,67, 1,641 95, 1,639 57

   c) 1,354 38, 1,946 33, 2,084 07

   d) 1,016 13, 1,030 67, 1,043 72

4. a) $f(1) = f'(1) = 0$, $f''(1) = 2$, racine de multiplicité 2

   b) $f(0) = 0$, $f'(0) = 1$, racine de multiplicité 1

   c) $f(0) = f'(0) = 0$, $f''(0) = 2$, racine de multiplicité 2

   d) $f(0) = 1$, 0 n'est pas une racine

5. a) $r = 0$, $g'(0) = 4$ (répulsif), $r = 3$, $g'(3) = -2$ (répulsif)

   b) $r = 0$, $g'(0) = \infty$ (répulsif), $r = 1$, $g'(1) = 1/2$ (attractif)

   c) $r = 0$, $g'(0) = 1$ (indéterminé)

   d) $r = \pm\sqrt{5}$, $g'(\pm\sqrt{5}) = 1 \mp 2\sqrt{5}$ (répulsifs)

6. a) L'algorithme de points fixes ne converge pas et oscille entre les valeurs ±2,236 067.

   b) $x_n \to 1{,}618\,0339, |e_n| \to 0, |e_n/e_{n-1}| \to 0{,}3090$, convergence linéaire

7. a) $x_n \to 2{,}690\,647, |e_n| \to 0, |e_n/e_{n-1}| \to 0, |e_n/(e_{n-1})^2| \to 0{,}5528$, convergence quadratique

   b) $x_n \to 1{,}872\,322, |e_n| \to 0, |e_n/e_{n-1}| \to 0, |e_n/(e_{n-1})^2| \to 0{,}5968$, convergence quadratique

   c) $x_n \to 1{,}134\,724, |e_n| \to 0, |e_n/e_{n-1}| \to 0, |e_n/(e_{n-1})^2| \to 2{,}44$, convergence quadratique

   d) $x_n \to 1{,}0, |e_n| \to 0, |e_n/e_{n-1}|$ ne tend pas vers 0, convergence linéaire

8. a) $x_0 = 3{,}0, x_1 = 4{,}0$: 2,826 086, 2,752 249, 2,694 940, 2,690 790, 2,690 647

   b) $x_0 = 2{,}0, x_1 = 3{,}0$: 1,913 390, 1,886 169, 1,872 646, 1,872 324, 1,872 322

   c) $x_0 = 1{,}5, x_1 = 2{,}5$: 1,461 637, 1,429 202, 1,263 943, 1,193 273, 1,149 644

   d) $x_0 = 1{,}2, x_1 = 2{,}2$: 1,198 062, 1,196 180, 1,135 400, 1,108 143, 1,081 331

9. On pose $g(x) = x - mf(x)/f'(x)$ et $f(x) = (x-r)^m h(x)$ avec $h(r) \neq 0$ et on montre que $g'(r) = 0$.

10. a) $g_1'(r_1) = -0{,}2294$ (attractif), $g_2'(r_1) = 0$ (attractif), $g_3'(r_1) = -0{,}099$ (attractif)

 b) La fonction $g_2(x)$ converge quadratiquement

 c) Oui, car il y a changement de signe

 d) Convergence quadratique vers 3,733 079

11. Il suffit de remarquer que $s^3 = 3 + s$. La méthode de Newton converge vers 1,671 699 88.

12. a) $|g'(x)| = |1 - 2\rho x| < 1$ pour $0 < \rho < \sqrt{2}/2$

 b) $g'(\sqrt{2}/4) = 0$, convergence quadratique

 c) $g'(3\sqrt{2}) > 1$, divergence

13. a) $g'(r) \simeq 0{,}51 < 1$

 b) $g'(1{,}365\,23) = 0{,}511\,96$

 c) $g'(1{,}365\,23) \neq 0$, convergence linéaire

14. a) La pente est fixée une fois pour toutes à $f'(x_0)$. Par conséquent, les droites sont toutes parallèles.

 b) On pose:

$$g(x) = x - \frac{f(x)}{f'(x_0)} = x - \frac{x^2 - 2}{2x_0}$$

 La condition de convergence est alors $|g'(r)| = g'(\sqrt{2})| < 1$. On obtient $\sqrt{2}/2 < x_0 < \infty$.

Réponses aux exercices du chapitre 3

1. a)

$$T = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \text{dét } T = 1$$

 b)

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \text{dét } P = -1$$

c)

$$M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \text{dét } M = 5$$

d)

$$T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 5 & 1 \end{bmatrix}, \text{dét } T = 1$$

2. a) $\vec{x} = [3 \;\; 2 \;\; 1]^T$, dét $A = 90$ b) $\vec{x} = [2 \;\; 2 \;\; -2]^T$, dét $A = -9$

3. a) Matrice augmentée triangularisée:

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 0 \\ 0 & -2 & 1 & 3 \\ 0 & 0 & 1/2 & 1/2 \end{array}\right]$$

dont la solution est $\vec{x} = [1 \;\; -1 \;\; 1]^T$. Le déterminant est -1.

b) Matrice augmentée triangularisée:

$$\left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & 0 & -5 & -15/2 & -35 \\ 0 & 0 & 0 & -9 & -18 \end{array}\right]$$

dont la solution est $\vec{x} = [3 \;\; -1 \;\; 4 \;\; 2]^T$. Le déterminant est -180.

4. Il suffit de construire les matrices correspondant à chacune des opérations élémentaires effectuées et de les multiplier par leur inverse pour obtenir une décomposition LU.

5. a) Matrice augmentée triangularisée (sans recherche de pivot):

$$\left[\begin{array}{ccc|c} 0{,}7290 & 0{,}8100 & 0{,}9000 & 0{,}6867 \\ 0{,}0000 & -0{,}1110 & -0{,}2350 & -0{,}1084 \\ 0{,}0000 & 0{,}0000 & 0{,}02640 & 0{,}008\,700 \end{array}\right]$$

dont la solution est $\vec{x} = [0{,}2251 \;\; 0{,}2790 \;\; 0{,}3295]^T$

b) Matrice augmentée triangularisée (avec recherche de pivot):

$$\left[\begin{array}{ccc|c} 1{,}331 & 1{,}210 & 1{,}100 & 1{,}000 \\ 0{,}0000 & 0{,}1473 & 0{,}2975 & 0{,}1390 \\ 0{,}0000 & 0{,}0000 & -0{,}010\,00 & -0{,}003\,280 \end{array}\right]$$

dont la solution est $\vec{x} = [0{,}2246 \;\; 0{,}2812 \;\; 0{,}3280]^T$

c) La solution en b) est plus précise.

6. a) Matrice augmentée triangularisée:

$$\left[\begin{array}{ccc|c} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & -2 \\ 0 & 0 & 0 & 2 \end{array}\right]$$

Le déterminant de la matrice est dét $A = (1)(1)(0) = 0$ et A est donc singulière.

b) La dernière équation signifie que $0 = 2$; il n'y a donc pas de solution.

7. a) Matrice augmentée triangularisée:

$$\left[\begin{array}{cc|c} 2 & -6\alpha & 3 \\ 0 & 9\alpha^2 - 1 & \beta - 9\alpha/2 \end{array}\right]$$

b) Le déterminant de la matrice est dét $A = (2)(9\alpha^2 - 1) = 18\alpha^2 - 2$.

c) $\alpha = \pm 1/3$

d) Si $\alpha = 1/3$, la matrice est singulière. Si $\beta = 1$, la dernière équation n'a pas de solution.

8. a) Décomposition LU sous forme compacte:

$$\left[\begin{array}{ccc} 1 & 2 & 1 \\ 2 & -2 & -1/2 \\ -1 & -1 & 1/2 \end{array}\right]$$

On obtient $\vec{y} = [0 \;\; -3/2 \;\; 1]^T$ et $\vec{x} = [1 \;\; -1 \;\; 1]^T$.

b) Décomposition LU sous forme compacte:

$$\left[\begin{array}{cccc} 1 & 2 & 1 & 4 \\ 2 & -4 & -1/2 & 5/4 \\ 4 & -6 & -5 & 3/2 \\ -3 & 7 & 19/2 & -9 \end{array}\right]$$

On obtient $\vec{y} = [13 \;\; -1/2 \;\; 7 \;\; 2]^T$ et $\vec{x} = [3 \;\; -1 \;\; 4 \;\; 2]^T$.

9. Décomposition LU sous forme compacte:

$$\begin{bmatrix} 4 & 2 & -1/4 \\ -2 & 7 & 3/14 \\ 1 & 0 & 25/4 \end{bmatrix}, O = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

On obtient $\vec{y} = [17/4 \;\; 37/14 \;\; 3]^T$ et $\vec{x} = [1 \;\; 2 \;\; 3]^T$.

10. a) dét $A = (1)(3)(6) = 18$

b) $||A||_\infty = \max(6\,, 27\,, 55) = 55$

c) Ces deux vecteurs sont les deux premières colonnes de A^{-1}. En résolvant $A\vec{x} = [0 \;\; 0 \;\; 1]^T$, on obtient la dernière colonne, qui est $[5/6 \;\; -2/3 \;\; 1/6]^T$.

d) $||A^{-1}||_\infty = \max(4{,}666\,53\,, 2{,}333\,29\,, 0{,}555\,55) = 4{,}666\,53$ et donc $\text{cond}A = 55 \times 4{,}666\,53 = 256{,}659$

11. a) dét $A = (2)(1)(2) = 4$

b) $\vec{y} = [-1 \;\; 18 \;\; 12]^T$ et $\vec{x} = [-4 \;\; -6 \;\; 12]^T$

c) On doit résoudre $A(A\vec{x}) = \vec{b}$. Il suffit de résoudre $A\vec{u} = \vec{b}$ et par la suite $A\vec{x} = \vec{u}$. Ces deux systèmes sont ensuite résolus au moyen de la décomposition LU de A. On trouve (voir a)) $\vec{u} = [-4 \;\; -6 \;\; 12]^T$ et ensuite $\vec{x} = [-2 \;\; 0 \;\; 1]^T$.

12. Décomposition LU sous forme compacte:

$$\begin{bmatrix} 0{,}500 & 2{,}00 & 4{,}00 \\ 0{,}333 & -0{,}416 & 2{,}72 \\ 0{,}250 & -0{,}300 & -0{,}0170 \end{bmatrix}, O = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

On obtient $\vec{y} = [16{,}0 \;\; -6{,}42 \;\; -181]^T$ et $\vec{x}^* = [-232 \;\; 486 \;\; -181]^T$. $||\vec{x} - \vec{x}^*||_\infty = \max(-4{,}92\,, 9{,}08\,, -3{,}30) = 9{,}08$ et l'erreur relative est donc $9{,}08/476{,}92 = 0{,}019$.

13.

$$A^{-1} = \begin{bmatrix} -83{,}077 & 64{,}615 & 0{,}461\,54 \\ 156{,}92 & -115{,}38 & -1{,}5384 \\ -57{,}692 & 41{,}538 & 1{,}1538 \end{bmatrix}$$

On a $||A||_\infty = \max(0{,}6166\,, 0{,}7833\,, 3{,}5) = 3{,}5$ et $||A^{-1}||_\infty = \max(148{,}15\,, 273{,}83\,, 100{,}38) = 273{,}83$. Enfin, $\text{cond}A = 3{,}5 \times 273{,}83 = 958{,}4$.

14. a) C'est le cercle de rayon 1 centré en l'origine.

    b) C'est le carré $[-1,1]^2$.

15. Non, puisque le déterminant peut s'annuler (si la matrice est singulière) sans que la matrice soit nulle.

16. Non, puisque A^{-1} n'existe pas.

17.

$$A^{-1} = \begin{bmatrix} -100 & 100 \\ 50{,}5 & -50 \end{bmatrix}$$

$||A||_\infty = 3{,}01$ et $||A^{-1}||_\infty = 200$. On a alors $\text{cond}A = 602$.

18.

$$A = \begin{bmatrix} 10^{-100} & 0 \\ 0 & 10^{-100} \end{bmatrix}$$

dont le déterminant est très petit (10^{-200}), mais dont le conditionnement est 1.

19. Le déterminant de cette matrice est $3{,}66 \times 10^{-12}$ et plusieurs logiciels la considèrent comme singulière. Elle est cependant inversible et son inverse est:

$$\begin{bmatrix} 52 & -300 & 1050 & -1400 & 630 \\ -300 & 4800 & -18\,900 & 26\,880 & -12\,600 \\ 1050 & -18\,900 & 79\,380 & -117\,600 & 56\,700 \\ -1400 & 26\,880 & -117\,600 & 179\,200 & -88\,200 \\ 630 & -12\,600 & 56\,700 & -88\,200 & 44\,100 \end{bmatrix}$$

20. La matrice est singulière et le système linéaire possède dans ce cas une infinité de solutions. Il est possible d'obtenir les 2 solutions proposées car numériquement, un système singulier devient souvent un système dont le déterminant est très petit et qui est donc mal conditionné.

21. a) $\vec{r}_1 = \vec{b} - A\vec{x}_1 = [-0{,}6 \ \ -0{,}600\,01]^T$, $\vec{r}_2 = \vec{b} - A\vec{x}_2 = [0{,}0 \ \ 0{,}0004]^T$, ce qui signifie que $||\vec{r}_1||_\infty = 0{,}600\,01$ et $||\vec{r}_2||_\infty = 0{,}0004$. De plus, $||\vec{x} - \vec{x}_1||_\infty = 0{,}1$ alors que $||\vec{x} - \vec{x}_2||_\infty = 4{,}0$. On en conclut que la solution approximative $\vec{x}_2$ est bien plus éloignée de la solution exacte,

mais donne un résidu beaucoup plus petit. Cela montre que la norme du résidu n'est pas toujours un bon indice de la qualité d'une solution.

b) La nouvelle solution est $[0 \;\; 1{,}2]^T$. Une petite perturbation du second membre entraîne une forte perturbation de la solution.

c) La matrice est mal conditionnée ($\text{cond}A = 120\,002$).

22.

$$J = \begin{bmatrix} 2x_1 - 10 & 2x_2 \\ x_2^2 + 1 & 2x_1x_2 - 10 \end{bmatrix}$$

Itération 1: $\vec{\delta x} = [0{,}8 \;\; 0{,}88]^T, \vec{x} = [0{,}8 \;\; 0{,}88]^T$

23.

$$J = \begin{bmatrix} 3 & x_3 \sin(x_2x_3) & x_2 \sin(x_2x_3) \\ 2x_1 & -162(x_2 + 0{,}1) & \cos x_3 \\ -x_2e^{-x_1x_2} & -x_1e^{-x_1x_2} & 20 \end{bmatrix}$$

Itération 1: $\vec{x}^1 = [0{,}499\,8697 \;\; 0{,}019\,4669 \;\; -0{,}521\,5205]^T$

Itération 2: $\vec{x}^2 = [0{,}500\,0142 \;\; 0{,}001\,5886 \;\; -0{,}523\,5569]^T$

Les itérations convergent vers $[0{,}5 \;\; 0 \;\; -0{,}523\,5987]^T$.

24. La matrice jacobienne est singulière en $[0 \;\; -0{,}2]^T$. Il faut alors amorcer la méthode de Newton à partir d'un autre point.

Réponses aux exercices du chapitre 4

1. Si x est un point fixe, alors $x = g(x)$ et $g(g(x)) = g(x) = x$. L'inverse est cependant faux.

2. Il suffit de calculer par exemple $g(x_1) = \lambda x_1(1 - x_1)$ et de montrer que l'on obtient x_2 (après simplification).

3. a) Le polynôme caractéristique est $\lambda^2 - 4\lambda + 3$. Les valeurs propres sont 1 et 3, et le rayon spectral est donc 3. La matrice est divergente.

   b) Le polynôme caractéristique est $(\lambda - 1/2)(\lambda - 1/3)(\lambda - 1/4)$. Les valeurs propres sont 1/4, 1/3 et 1/2, et le rayon spectral est donc 1/2. La matrice est convergente.

4. La convergence de la méthode de Jacobi dépend du rayon spectral de la matrice $T_J = -D^{-1}(T_i + T_s)$ et non du rayon spectral de la matrice A elle-même. Dans ce cas:

$$T_j = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix}$$

et son rayon spectral est $1/2$.

5. a) Seul $(1, 1)$ est un point fixe.

   b) Il est attractif car $\rho(J(1, 1)) = \sqrt{2}/2 < 1$.

   c) Les 5 premières itérations donnent: (1,4142 , 0), (1,4142 , 1,1892), (0,7653 , 1,1892), (0,7654 , 0,8749) et (1,1111 , 0,8749).

6. a) Il faut démontrer que $|g_1'(x_1)| < 1$, où $g_1(x) = g(g(x))$. Le résultat vient immédiatement de la règle de dérivation en chaîne.

   b) Si $\{x_1, x_2, x_3, \cdots, x_n\}$ est un n-cycle, alors il est attractif si:

$$\prod_{i=1}^{n} |g'(x_i)| < 1$$

7. $\{-1, 0\}$ est un 2-cycle attractif.

8. Il suffit de montrer que $T(2/7) = 4/7$, $T(4/7) = 6/7$ et $T(6/7) = 2/7$, et que $T(2/9) = 4/9$, $T(4/9) = 8/9$ et $T(8/9) = 2/9$. Ces 3-cycles sont tous deux répulsifs, car $|T'(2/7)T'(4/7)T'(6/7)| = 8$ et $|T'(2/9)T'(4/9)T'(8/9)| = 8$.

9. Il suffit d'expliciter le terme $T\vec{x} + \vec{c}$ et de constater que la matrice jacobienne est T. Cette méthode de points fixes convergera si $\rho(T) < 1$. On remarque de plus que les méthodes de Jacobi et de Gauss-Seidel sont de cette forme.

10. On peut prendre par exemple la matrice:

$$\begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{bmatrix}$$

pour laquelle la matrice T_J est:

$$\begin{bmatrix} 0 & -1/3 & -1/3 \\ -1/3 & 0 & -1/3 \\ -1/3 & -1/3 & 0 \end{bmatrix}$$

On a alors $||T_J||_\infty = 2/3$, qui est inférieur à 1.

11. a) Dès la première équation, $a_{11} = 0$ et la méthode de Jacobi s'arrête.

    b) On réordonne les équations de telle sorte que la nouvelle matrice soit à diagonale strictement dominante: E_3, E_4, E_1, E_2.

12. Les itérations de la méthode de Jacobi donnent:

n	x_1^n	x_2^n	x_3^n
1	1,444 444	1,800 000	−1,222 222
2	1,980 247	1,844 444	−0,982 7160
3	1,963 512	1,999 506	−1,032 373
4	2,003 487	1,986 228	−0,996 0555
5	1,996 501	2,001 486	−1,003 448

Avec la méthode de Gauss-Seidel, on obtient:

n	x_1^n	x_2^n	x_3^n
1	1,444 444	2,088 889$E+01$	−0,918 5185
2	2,010 700	2,018 436$E+01$	−0,997 0919
3	2,003 774	2,001 336$E+01$	−1,000 122
4	2,000 311	2,000 038$E+01$	−1,000 026
5	2,000 011	1,999 997$E+01$	−1,000 002

La méthode de Gauss-Seidel converge plus rapidement vers $[2\ 2\ -1]^T$.

Réponses aux exercices du chapitre 5

1. Cette affirmation est vraie en général. Cependant, dans certains cas, il est possible de construire ce polynôme. Par exemple, si on choisit 3 points sur une droite, on peut construire le polynôme de degré 1 (la droite) passant par ces 3 points.

2. On doit résoudre le système:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 6 \\ 12 \end{bmatrix}$$

dont la solution est $[0 \ \ 1 \ \ 1]^T$. Le polynôme est donc $p_2(x) = x + x^2$.

3.

$$\text{a) } p_2(x) = 0\frac{(x-1)(x-2)}{(0-1)(0-2)} + 2\frac{(x-0)(x-2)}{(1-0)(1-2)} + 36\frac{(x-0)(x-1)}{(2-0)(2-1)}$$

$$\begin{aligned} \text{b) } p_3(x) &= 0\frac{(x-1)(x-2)(x-3)}{(0-1)(0-2)(0-3)} + 2\frac{(x-0)(x-2)(x-3)}{(1-0)(1-2)(1-3)} \\ &+ 36\frac{(x-0)(x-1)(x-3)}{(2-0)(2-1)(2-3)} + 252\frac{(x-0)(x-1)(x-2)}{(3-0)(3-1)(3-2)} \end{aligned}$$

$$\begin{aligned} \text{c) } E_2(x) &= \frac{f^{(3)}(\xi)(x)(x-1)(x-2)}{3!}, \ \ \xi \in [0,2] \\ E_3(x) &= \frac{f^{(4)}(\xi)(x)(x-1)(x-2)(x-3)}{4!}, \ \ \xi \in [0,3] \end{aligned}$$

d) $p_2(1{,}5) = 15{,}0, \ \ p_3(1{,}5) = 5{,}625$

4. a) $p_2(x) = 2x + 16x(x-1)$

b) $p_3(x) = p_2(x) + 25x(x-1)(x-2)$

c) Les expressions analytiques des erreurs sont les mêmes qu'à l'exercice précédent. Cependant, on peut estimer la valeur de ces erreurs:

$$E_2(x) \simeq 25x(x-1)(x-2), \ \ E_3(x) \simeq 10x(x-1)(x-2)(x-3)$$

d) Mêmes réponses qu'au numéro précédent. De plus, $E_2(1{,}5) \simeq -9{,}375$ et $E_3(1{,}5) \simeq 5{,}625$.

5. a) Le système linéaire est:

$$\begin{bmatrix} 2 & 1/2 & 0 \\ 1/2 & 2 & 1/2 \\ 0 & 1/2 & 2 \end{bmatrix} \begin{bmatrix} f_1'' \\ f_2'' \\ f_3'' \end{bmatrix} = \begin{bmatrix} 96 \\ 546 \\ 1716 \end{bmatrix}$$

b) La solution est $[34{,}7143 \ \ 53{,}1428 \ \ 844{,}714]^T$.

c) L'abscisse 1,5 est dans le deuxième intervalle ($i = 2$). L'équation de la spline est:

$$\begin{aligned} p(x) &= -34{,}7143(x-2)^3/6 + 53{,}1428(x-1)^3/6 \\ &\quad - (2 - 34{,}7143/6)(x-2) + (36 - 53{,}1428/6)(x-1) \end{aligned}$$

ce qui signifie que $p(1{,}5) = 13{,}5089$.

6. a) On prend les points dont les abscisses sont les plus rapprochées de 4,5 (il y a deux possibilités). En prenant les points d'abscisse 5, 3,5 et 7,0, on trouve:

$$p_2(x) = 1{,}6094 + 0{,}237\,733(x-5) - 0{,}019\,8523(x-5)(x-3{,}5)$$

qui prend la valeur 1,500 459 65 en $x = 4{,}5$. L'expression analytique du terme d'erreur est:

$$E_2(x) = \frac{f^{(3)}(\xi)(x-5)(x-3{,}5)(x-7)}{3!} \quad \text{avec } \xi \in [3{,}5\,,\,7]$$

b)

$$\begin{aligned} p_2(x) &= 1{,}6094\frac{(x-3{,}5)(x-7)}{(5-3{,}5)(5-7)} + 1{,}2528\frac{(x-5)(x-7)}{(3{,}5-5)(3{,}5-7)} \\ &\quad + 1{,}9459\frac{(x-5)(x-3{,}5)}{(7-5)(7-3{,}5)} \end{aligned}$$

c) $E_2(x) \simeq 0{,}005(x-5)(x-3{,}5)(x-7)$, de telle sorte que $E_2(4{,}5) \simeq 0{,}006\,25$

d) Non. Il faut utiliser la méthode de Newton.

e) Les deux méthodes donnent le même polynôme, mais exprimé différemment. Elles ont le même terme d'erreur, mais seule la méthode de Newton peut fournir une approximation de l'erreur.

7. On donne le tableau des valeurs des différents polynômes en fonction du degré ainsi que l'approximation de l'erreur commise.

n	$p_n(1{,}05) =$	$E_n(1{,}05) \simeq$
1	0,852 835	$0{,}298\,75 \times 10^{-3}$
2	0,853 133 75	$0{,}375 \times 10^{-5}$
3	0,853 1375	$-0{,}1562 \times 10^{-5}$

On constate que l'erreur absolue est inférieure à $0{,}2 \times 10^{-5}$ pour le polynôme de degré 3.

8.

$$\begin{aligned} f[x_i, x_{i+1}] &= \frac{f(x_{i+1}) - f(x_i)}{h} \\ f[x_i, x_{i+1}, x_{i+2}] &= \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{2h^2} \\ f[x_i, x_{i+1}, x_{i+2}, x_{i+3}] &= \frac{f(x_{i+3}) - 3f(x_{i+2}) + 3f(x_{i+1}) - f(x_i)}{3!h^3} \end{aligned}$$

9. a) $2f_1'' = 18$, d'où $f_1'' = 9$

   b) $p_1(x) = (1/2)(3x^3 - x)$, d'où $p_1(1/2) = -0{,}0625$

   c) Pour la spline naturelle, on utilise l'approximation $f_0'' = f_2'' = 0$. Cependant, la fonction $f(x) = x^3$ a comme dérivée seconde $f''(x) = 6x$ qui ne s'annule pas en $x = 2$, d'où l'erreur.

10. a) La nouvelle équation s'écrit:

$$f_0''/2 + 2f_1'' + f_2''/2 = 18$$

   ce qui devient $2f_1'' = 18 - 6 = 12$, d'où $f_1'' = 6$.

   b) L'équation de la spline est alors $p_2(x) = x^3$.

   c) L'approximation est exacte.

11. a) Il y a 4 conditions. Il faut donc un polynôme de degré 3.

    b) $p_3(x) = x^2(2{,}7 - 1{,}7x)$

12. a) Il suffit de calculer la table de différences divisées.

    b) La fonction $f(x)$ inconnue est un polynôme de degré 2 dont l'équation est $p_2(x) = x^2 - 2x + 3$.

13. a)

$$\begin{bmatrix} 0 & 2 & 5 & 1 & 0 \\ 2 & 0 & 3 & 1 & 2 \\ 5 & 3 & 0 & 1 & 5 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 2 & 5 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 7 \\ 0 \\ 0 \end{bmatrix}$$

b) $u(x) = 0{,}15|x - 0| - 0{,}25|x - 2| + 0{,}1|x - 5| + 1{,}0 + 1{,}2x$ et $u(3) = 5$

c) Il suffit d'effectuer une interpolation linéaire entre les deux derniers points de la table et on obtient le même résultat en $x = 3$.

14. a)

$$\begin{bmatrix} 0 & 8 & 125 & 1 & 0 \\ 8 & 0 & 27 & 1 & 2 \\ 125 & 27 & 0 & 1 & 5 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 2 & 5 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 7 \\ 0 \\ 0 \end{bmatrix}$$

b) $u(x) = -0{,}0125|x - 0|^3 + 0{,}020\,833|x - 2|^3 - 0{,}008\,333|x - 5|^3 + 1{,}875 + 1{,}225x$ et $u(3) = 5{,}1667$

15.

$$\begin{bmatrix} 0 & 4 & 25 & 1 & 0 \\ 4 & 0 & 9 & 1 & 2 \\ 25 & 9 & 0 & 1 & 5 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 2 & 5 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 7 \\ 0 \\ 0 \end{bmatrix}$$

Le déterminant de cette matrice est nul.

16.

$$\begin{bmatrix} 0 & 0 & 0 & 0{,}693\,147 & 1 & 1 & 1 \\ 0 & 0 & 0{,}693\,147 & 0 & 1 & 2 & 1 \\ 0 & 0{,}693\,147 & 0 & 0 & 1 & 1 & 2 \\ 0{,}693\,1479 & 0 & 0 & 0 & 1 & 2 & 2 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 2 & 0 & 0 & 0 \\ 1 & 1 & 2 & 2 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 4 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

En résolvant le système, on trouve $\alpha_1 = 0{,}360\,674$, $\alpha_2 = -0{,}360\,674$, $\alpha_3 = -0{,}360\,674$, $\alpha_4 = 0{,}360\,674$, $a_1 = -2{,}25$, $a_2 = 1{,}5$ et $a_3 = 1{,}5$. La fonction:

$$u(x_1, x_2) = \sum_{j=1}^{4} \alpha_j ||\vec{x} - \vec{x}^j||_e^2 \ln ||\vec{x} - \vec{x}^j||_e + a_1 + a_2 x_1 + a_3 x_2$$

évaluée en $(3/2, 3/2)$, vaut précisément $9/4$.

17. Suivant la formule 5.48, on obtient $t_1 = 0$, $t_2 = 1$, $t_3 = 2$, $t_4 = 3$ et $t_5 = 4$. Les systèmes à résoudre sont:

$$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 1 & 0 \\ 1 & 0 & 1 & 2 & 3 & 1 & 1 \\ 2 & 1 & 0 & 1 & 2 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 & 1 & 3 \\ 4 & 3 & 2 & 1 & 0 & 1 & 4 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 3 & 4 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

En résolvant ces systèmes, on trouve $[1/2 \ \ -1/2 \ \ -1/2 \ \ 1/2 \ \ 0 \ \ 0 \ \ 0]^T$ et $[0 \ \ 1/2 \ \ -1/2 \ \ -1/2 \ \ 1/2 \ \ 0 \ \ 0]^T$, ce qui signifie que:

$$\gamma_1(t) = \frac{1}{2}|t| - \frac{1}{2}|t-1| - \frac{1}{2}|t-2| + \frac{1}{2}|t-3|$$

et

$$\gamma_2(t) = \frac{1}{2}|t-1| - \frac{1}{2}|t-2| - \frac{1}{2}|t-3| + \frac{1}{2}|t-4|$$

qui est l'équation paramétrique du carré.

Réponses aux exercices du chapitre 6

1. Voir page 298

2. Voir page 298

3. Différences avant: $f'(0) \simeq 0{,}999\,135$ pour $h = 0{,}05$, $f'(0) \simeq 0{,}999\,788$ pour $h = 0{,}025$. Richardson: $(4 \times 0{,}999\,788 - 0{,}999\,135)/3 = 1{,}000\,005\,667$

   Différences arrière: $f'(0) \simeq 0{,}999\,1972$ pour $h = 0{,}05$, $f'(0) \simeq 0{,}999\,7955$ pour $h = 0{,}025$. Richardson: $(4 \times 0{,}999\,7955 - 0{,}999\,1972)/3 = 0{,}999\,994\,96$

4.
$$\frac{f(x+h) - f(x-h)}{2h} = f'(x) + \frac{h^2 f'''(x)}{3!} + \frac{h^4 f^{(5)}(x)}{5!} + \frac{h^6 f^{(7)}(x)}{7!} + \cdots$$

   L'extrapolation de Richardson permet de gagner 2 ordres de précision.

5. Il faut calculer le polynôme de degré 4 passant par les points $(x-2h, f(x-2h))$, $(x-h, f(x-h))$, $(x, f(x))$, $(x+h, f(x+h))$, $(x+2h, f(x+2h))$, et le dériver 2 fois.

6. Il suffit d'intégrer respectivement $p_1(x)$ sur l'intervalle $[x_0, x_1]$, $p_2(x)$ sur l'intervalle $[x_0, x_2]$ et $p_3(x)$ sur l'intervalle $[x_0, x_3]$.

7. Trapèzes: 1,727 221 905 pour 4 intervalles (h = 0,25) et 1,720 518 592 pour 8 intervalles (h = 0,125). Richardson: $(4 \times 1{,}720\,518\,592 - 1{,}727\,221\,905)/3 = 1{,}718\,284\,154$ (ordre 4). Les erreurs respectives sont $0{,}894 \times 10^{-2}, 0{,}223 \times 10^{-2}$ et $0{,}2326 \times 10^{-5}$.

8. Simpson 1/8: 1,718 318 843 pour 4 intervalles (h = 0,25) et 1,718 284 154 pour 8 intervalles (h = 0,125). Richardson: $(2^4 \times 1{,}718\,284\,154 - 1{,}718\,318\,843)/15 = 1{,}718\,281\,843$ (ordre 6). Les erreurs respectives sont $0{,}37 \times 10^{-4}, 0{,}2326 \times 10^{-5}$ et $0{,}15 \times 10^{-7}$.

9. Simpson 3/8 ($h = 4/3$): 17,327 866 29 avec une erreur absolue de $0{,}54 \times 10^{-2}$

10. Boole ($h = \pi/32$): 0,881 374 32 avec une erreur absolue de $0{,}733 \times 10^{-6}$

11.

0,785 398 164	0,916 297 857	0,968 361 509	(ordre 2)
0,959 931 089	0,985 716 059		(ordre 4)
0,987 435 057			(ordre 6)

12. Il suffit de développer l'expression.

13. a) ln 2

    b) Il faut se rendre jusqu'à $T_{1,4}$, pour lequel le rapport est 0,004.

    c)

0,75	0,708 333 33	0,697 023 809	0,694 121 85	(ordre 2)
0,694 444 444	0,693 253 968	0,693 154 53		(ordre 4)
0,693 174 603	0,693 147 901			(ordre 6)
0,693 147 901				(ordre 8)

    d) La deuxième ligne du tableau correspond à la méthode de Simpson 1/3 avec 2, 4 et 8 intervalles.

14. Il faut utiliser les méthodes de Gauss, car la fonction $\ln x$ n'est pas définie en $x = 0$. Les formules à 2, à 3 et à 5 points donnent respectivement les approximations $-0{,}405\,465$, $-0{,}509\,050\,405$ et $-0{,}571\,707\,615$. La valeur exacte est $-0{,}613\,705\,639$.

15. La formule à 3 points est exacte pour les polynômes de degré 5.

16. a) $a = b = 3h/2$

    b) 1,25

    c) 1,425

    d) Respectivement 1 et 2 chiffres significatifs

17.

$$\frac{\text{Erreur}(h = 0{,}2)}{\text{Erreur}(h = 0{,}1)} = \frac{0{,}009\,872}{0{,}001\,234} = 7{,}99 \simeq 2^3$$

La méthode est donc d'ordre 3.

18. a) Le terme de droite devient:

$$hf(x_0) + \frac{h^2 f'(x_0)}{2} + \frac{h^3 f''(x_0)}{6} + \frac{h^4 f'''(x_0)}{27} + \frac{h^5 f''''(x_0)}{162} + O(h^6)$$

b) Le terme de gauche devient, après intégration:

$$hf(x_0) + \frac{h^2 f'(x_0)}{2} + \frac{h^3 f''(x_0)}{6} + \frac{h^4 f'''(x_0)}{24} + \frac{h^5 f''''(x_0)}{120} + O(h^6)$$

c) Le premier terme de l'erreur est:

$$h^4 f'''(x_0)\left(\frac{1}{24} - \frac{1}{27}\right)$$

et la méthode est d'ordre 4.

d) Degré 2

Réponses aux exercices du chapitre 7

1. a) Euler: $y_1 = 2$, $y_2 = 2{,}009\,0929$, $y_3 = 2{,}027\,202\,49$

   Euler modifiée: $y_1 = 2{,}004\,546\,487, y_2 = 2{,}018\,118\,919, y_3 = 2{,}040\,539\,939$

Runge-Kutta d'ordre 4: $y_1 = 2{,}004\,541\,741$, $y_2 = 2{,}018\,109\,47$, $y_3 = 2{,}040\,526\,45$

b) Euler: $y_1 = 0{,}2$, $y_2 = 0{,}425$, $y_3 = 0{,}687\,0625$

Euler modifiée: $y_1 = 0{,}2125$, $y_2 = 0{,}456\,850\,69$, $y_3 = 0{,}749\,830\,45$

Runge-Kutta d'ordre 4: $y_1 = 0{,}211\,7831$, $y_2 = 0{,}455\,527\,18$, $y_3 = 0{,}748\,199$

c) Euler: $y_1 = 2{,}2$, $y_2 = 2{,}443\,1376$, $y_3 = 2{,}741\,543$

Euler modifiée: $y_1 = 2{,}221\,5688$, $y_2 = 2{,}494\,994$, $y_3 = 2{,}836\,326$

Runge-Kutta d'ordre 4: $y_1 = 2{,}221\,8007$, $y_2 = 2{,}495\,651$, $y_3 = 2{,}837\,7328$

2. a) Pour $h = 0{,}1$, $y(0{,}3) \simeq y_3 = 3{,}170\,000\,1557$ avec une erreur absolue de $0{,}001\,977$

   b) Pour $h = 0{,}05$, $y(0{,}3) \simeq y_6 = 3{,}171\,450\,217$ avec une erreur absolue de $0{,}000\,527$

   c) Le ratio des erreurs est de $3{,}75 \simeq 2^2$, ce qui confirme que la méthode est d'ordre 2.

   d) Richardson: $(2^2 \times 3{,}171\,450\,217 - 3{,}170\,000\,1557)/3 = 3{,}171\,933\,572$

3. a) Pour $h = 0{,}1$, $y(0{,}3) \simeq y_3 = 3{,}171\,976\,0094$ avec une erreur absolue de $0{,}1599 \times 10^{-5}$

   b) Pour $h = 0{,}05$, $y(0{,}3) \simeq y_6 = 3{,}171\,977\,5025$ avec une erreur absolue de $0{,}83 \times 10^{-7}$

   c) Le ratio des erreurs est de $19{,}26 \simeq 2^4$, ce qui confirme que la méthode est d'ordre 4.

   d) Richardson: $(2^4 \times 3{,}171\,977\,5025 - 3{,}171\,976\,0094)/15 = 3{,}171\,977\,601$

4. Il suffit d'utiliser la définition de l'erreur de troncature locale et de faire les développements de Taylor appropriés.

5. $y_1^1 = 2{,}331\,733$, $y_2^1 = 1{,}321\,041$, $y_1^2 = 2{,}734\,468$, $y_2^2 = 1{,}688\,708$

6. a)

$$\begin{array}{rcll} y_1'(t) & = & y_2(t) & (y_1(0) = 2) \\ y_2'(t) & = & y_3(t) & (y_2(0) = 2) \\ y_3'(t) & = & y_3(t) + y_2(t) - y_1(t) + 1 & (y_3(0) = 1) \end{array}$$

b)

$$\begin{array}{rcll} y_1'(t) & = & y_2(t) & (y_1(1) = 0) \\ y_2'(t) & = & (y_1(t))^2 + t^2 + 1 & (y_2(1) = 2) \end{array}$$

c)

$$\begin{array}{rcll} y_1'(t) & = & y_2(t) & (y_1(0) = 2) \\ y_2'(t) & = & y_3(t) & (y_2(0) = 1) \\ y_3'(t) & = & y_4(t) & (y_3(0) = 0) \\ y_4'(t) & = & e^t y_3(t) + (y_4(t))^3 & (y_4(0) = 4) \end{array}$$

7. a)

$$\begin{array}{rcl} y_1''(x) & = & -y_1'(x) + 2y_1(x) - 16 \\ y_1(0) & = & -7 \\ y_1'(0) & = & 0 \end{array}$$

$$\begin{array}{rcl} y_2''(x) & = & -y_2'(x) + 2y_2(x) \\ y_2(0) & = & 0 \\ y_2'(0) & = & 1 \end{array}$$

b)

$$\begin{array}{rcll} u_1'(t) & = & u_2(t) & (u_1(0) = -7) \\ u_2'(t) & = & -u_2(t) + 2u_1(t) - 16 & (u_2(0) = 0) \end{array}$$

$$\begin{array}{rcll} v_1'(t) & = & v_2(t) & (v_1(0) = 0) \\ v_2'(t) & = & -v_2(t) + 2v_1(t) & (v_2(0) = 1) \end{array}$$

c) On résout les 2 systèmes et on pose:

$$y(x) = y_1(x) + \frac{12{,}0855 - y_1(3)}{y_2(3)} y_2(x) = u_1(x) + \frac{12{,}0855 - u_1(3)}{v_1(3)} v_1(x)$$

8. Il suffit de prouver que $y(x)$ est bien une solution de l'équation différentielle et que $y(a) = y_a$ et $y'(b) = y_b'$.

9. Il suffit de prouver que $y(x)$ est bien une solution de l'équation différentielle et que $y'(a) = y_a'$ et $y(b) = y_b$.

10. Prédiction-correction d'ordre 2: la première valeur a été obtenue à l'exercice 1a) à l'aide de la méthode de Runge-Kutta d'ordre 4.

t	y_n^p	y_n
0,1		2,004 5417
0,2	2,018 1527	2,018 0947
0,3	2,040 6062	2,040 4857
0,4	2,071 5964	2,071 4053

Prédiction-correction d'ordre 4: les 3 premières valeurs ont été obtenues à l'exercice 1a) à l'aide de la méthode de Runge-Kutta d'ordre 4.

t	y_n^p	y_n
0,1		2,004 5417
0,2		2,018 1095
0,3		2,040 5264
0,4	2,071 4899	2,071 4842
0,5	2,110 5338	2,110 5267
0,6	2,157 0371	2,157 0304

11. Le système est de dimension 4 et prend la forme:

$$\begin{bmatrix} 4{,}88 & -2{,}0 & 0{,}0 & 0{,}0 \\ -2{,}0 & 4{,}48 & -2{,}0 & 0{,}0 \\ 0{,}0 & -2{,}0 & 4{,}336\,67 & -2{,}0 \\ 0{,}0 & 0{,}0 & -2{,}0 & 4{,}28 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0{,}176 \\ 0{,}192 \\ 0{,}208 \\ 4{,}224 \end{bmatrix}$$

dont la solution est $[0{,}290\,236 \quad 0{,}620\,176 \quad 1{,}002\,959 \quad 1{,}455\,588]^T$.

12. On pose $a_0(x) = -(x+2)$, $a_1(x) = 1 + 2/x$ et $a_2(x) = 0$, et on obtient le système:

$$-y_{i-1} + (2 + ha_1(x_i) + h^2 a_2(x_i))y_i + (-1 + ha_2(x_i))y_{i+1} = -h^2 a_0(x_i)$$

pour $i = 1, 2, 3, \ldots, (n-1)$. La première équation ($i = 1$) fait intervenir $y_0 = y_a$ et le terme correspondant est envoyé à droite. De même, la dernière équation ($i = n-1$) utilise $y_n = y_b$. L'ordre de cette méthode de différences finies est 1.

Bibliographie

1. Bourdeau, M. et Gélinas, J., *Analyse numérique élémentaire*, Chicoutimi, Gaëtan Morin éditeur, 1982.

2. Burden, R.L. et Faires, J.D., *Numerical Analysis*, 5e édition, Boston, PWS Kent, 1993.

3. Carreau, P.J., De Kee, D. et Chhabra, R.P., *Rheology of Polymeric Systems: Principles and Applications.* À paraître.

4. Chapra, C.S. et Canale, R.P., *Numerical Methods for Engineers*, 2e édition, New York, McGraw-Hill, 1985.

5. Cheney, W. et Kincaid, D., *Numerical Mathematics and Computing*, 3e édition, Pacific Grove, Brooks/Cole, 1994.

6. Conte, S.D. et de Boor, C., *Elementary Numerical Analysis, An Algorithmic Approach*, 3e édition, New York, McGraw-Hill, 1980.

7. Derrick, W.R. et Grossman, S.I., *Introduction to Differential Equations*, 3e édition, St-Paul, West Publishing Company, 1987.

8. Derrida, B., Gervois, A. et Pomeau, Y., ≪Universal Metric Properties of Bifurcations and Endomorphisms ≫, *Journal of Physics A*, vol. 12, n°3, p. 269-296, 1979.

9. Duchon, J., ≪Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces ≫, *RAIRO, Analyse numérique*, vol. 10, p. 5-12, 1976.

10. Feigenbaum, M., ≪Universal Behavior in Nonlinear Systems ≫, *Physica 5D*, p.16-39, 1983.

11. Fortin, M. et Pierre, R., *Analyse numérique MAT-10427*, Québec, Université Laval, 1993.

12. Gerald, C.F. et Wheatly, P.O., *Applied Numerical Analysis*, 4e édition, Reading, Addison-Wesley, 1989.

13. Gulick, D., *Encounters with Chaos*, New York, McGraw-Hill, 1992.

14. Institute for Electric and Electronic Engineers, *IEEE Recommended Practices and Requirements for Harmonic Control in Electrical Power Systems*, Norme Std 519-1992, New York, IEEE, 1993.

15. Kreyszig, E., *Advanced Engineering Mathematics*, 6e édition, New York, Wiley, 1988.

16. Krige, D.G., ≪A Statistical Method for Mine Variation Problems in the Witwatersrand≫, *Journal of Chemistry and Metallurgy of the Mining Society of South Africa*, vol. 52, p. 119-139, 1951.

17. Lorenz, E.N., ≪Deterministic Nonperiodic Flow≫, *Journal of Atmospheric Sciences*, vol. 20, p. 130-141, 1963.

18. Mandelbrot, B., *The Fractal Geometry of Nature*, San Francisco, W.H. Freeman and Co., 1982.

19. Matheron, G., ≪The Intrinsec Ramdom Functions and their Applications≫, *Advances in Applied Probability*, vol. 5, p. 439-468, 1973.

20. Simmons, G.S., *Differential Equations with Applications and Historical Notes*, New York, McGraw-Hill, 1972.

21. Strang, G., *Introduction to Applied Mathematics*, Wellesley, Wellesley-Cambridge Press, 1986.

22. Thomas, B.T. Jr. et Finney, R.L., *Calculus and Analytic Geometry*, 8e édition, Reading, Addison-Wesley, 1992.

23. Trochu, F., ≪A Contouring Program Based on Dual Kriging Interpolation≫, *Engineering with Computers*, vol. 9, p. 160-177, 1993.

24. Reddy, J.N., *An Introduction to the Finite Element Method*, 2e édition, New York, McGraw-Hill, 1993.

25. Scheid, F., *Numerical Analysis*, Schaum's Outline Series, New York, McGraw-Hill, 1968.

26. Varga, R., *Matrix Iterative Analysis*, Englewood Cliffs, Prentice-Hall, 1962.

Index